University of New Orleans

ScholarWorks@UNO

University of New Orleans Theses and Dissertations

Dissertations and Theses

Summer 8-2-2012

Evolution of Nuclear Integrations of the Mitochondrial Genome in Great Apes and their Potential as Molecular Markers

Ivan D. Soto-Calderon
University of New Orleans, ivandariosoto@hotmail.com

Recommended Citation

Soto-Calderon, Ivan D., "Evolution of Nuclear Integrations of the Mitochondrial Genome in Great Apes and their Potential as Molecular Markers" (2012). *University of New Orleans Theses and Dissertations*. 1510.
https://scholarworks.uno.edu/td/1510

Evolution of Nuclear Integrations of the Mitochondrial Genome in Great Apes and their Potential as Molecular Markers

A Dissertation

Submitted to the Graduate Faculty of the
University of New Orleans
in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
In
Conservation Biology

By

Iván Darío Soto Calderón

B.Sc, University of Antioquia, 1999.
M.Sc. University of Antioquia, 2003.

August 2012

ACKNOWLEDGEMENTS

Many people have guided me along this process, especially my parents who gave me the opportunity for me to familiarize myself with the country side and enjoy wildlife in my childhood. Leydy Laura Marín encouraged me and gave me her unconditional support during the early years of my doctoral studies, which was essential to stand firm during the hardest moments. My dear friend Daniel Ortiz-Barrientos was always positive and helped me in making the best decisions. Also, María Isabel Montes was a loving and constant company during the last part of my studies. Many people surrounded me after Katrina, helping in one or other way. Many thanks to my friends David Quiceno and Milena Zuluaga, as well as Kyle Harms and his laboratory members, Heather Passmore and Paul Gagnon and, Nathan and Heather Jackson in Baton Rouge; also, thanks to Adriana Obando and Andrés Ramírez, who hosted me in Houston after the hurricane.

This work would not have been possible without the constant support and the insightful suggestions of my major professor Nicola Anthony, who trusted me, gave me the opportunity to join her laboratory, allowed me to contribute to her own research projects with butterflies and duikers and provided me with the means to participate in scientific meetings and field work. I would like to recognize the wonderful undergraduate students Eun-Joo Lee and Jonathan Clark for their enthusiasm and tremendous help. I am also grateful to all the members of my thesis committee, Mary Clancy, Charles Bell, Steven Johnson, Prescott Deininger and Michael Jensen-Seaman, who always gave me friendly and constructive advice. Mike also adopted me as a member of his laboratory during my stay at Duquesne University, where I had the opportunity to improve my skills in cell culture and quantitative PCR thanks to the knowledgeable and amazing graduate student Sarah Carnahan-Craig.

I would also like to thank Dr. Richard Bergl from North Carolina Zoo, Naida Loskutoff from Omaha's Henry Doorly Zoo, Betsy Dresser from the Audubon Center for Research of Endangered Species and Barry Bavister from UNO for providing critical DNA samples from gorillas captive in USA institutions. Likewise, access to fecal samples of wild ranging gorillas was made possible thanks to support from Mary Katherine Gonder at the State University of New York at Albany, the unforgettable field work experience with Sabrina Locatelli in Cameroon and special collaboration of people from Wildlife Conservation Society – Cameroon. This work was also possible thanks to the technical and logistic support of UNO staff, particularly Robin

Rowe, Yvette Stilley, Teresa Howell, Jeanne Guimond and Chancy Hence, as well as faculty members Bernard Rees, Wendy Schluchter and Jerry Howard.

Finally, funds needed to conduct this research and cover my expenses were generously provided by UNO as well as the National Institutes of Health and National Geographic Society through grants to my major professor Nicola Anthony.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABSTRACT

The mitochondrial control region (MCR) has played an important role as a population genetic marker in many taxa but sequencing of complete eukaryotic genomes has revealed that nuclear integrations of mitochondrial DNA (numts) are abundant and widespread across many taxa. If left undetected, numts can inflate mitochondrial diversity and mislead interpretation of phylogenetic relationships. Comparative analyses of complete genomes in humans, orangutans and chimpanzees, and preliminary studies in gorillas have revealed high numt prevalence in great apes, but rigorous comparative analyses across taxa have been lacking.

The present study aimed to systematically compare the evolutionary dynamics of MCR numts in great apes. Firstly, an inventory numts derived from the region containing the MCR subdomains was carried out by genomic BLAST searches. Secondly, presence/absence of each candidate numt was determined in great ape taxa to estimate numt insertion rate. Thirdly, alternative mechanisms of numt insertion, either through direct mitochondrial integration or post-insertional duplications, were also assessed. Fourthly, the effect of nuclear and mitochondrial environment on patterns of nucleotide composition and substitution was assessed through sequence comparisons of nuclear and mitochondrial paralogous sequences. Finally, numts in the gorilla genome were identified through two experimental methods and their use as polymorphic genetic markers was then evaluated in a sample of captive gorillas from U.S. zoos.

A deficit of MCR numts covering two particular mitochondrial subdomains was detected in all three apes examined, and is largely attributed to rapid loss of mitochondrial and nuclear sequence identity in the mitochondrial genome. Insertion rates have varied during the great ape evolution and exhibit substantial differences even between related taxa. The most likely mechanism of numt insertion is direct mitochondrial integration through Non-Homologous-End-Joining Repair. Transition/transversion ratios differed significantly between both mitochondrial and nuclear sequences and between numts from coding and non-coding mitochondrial regions. A previously documented upward bias in the GC content of the primate mitochondrial genome was confirmed and the extent of this bias relative to the corresponding numt sequences increased with numt age. Five gorilla-specific numts were isolated, including three exhibiting insertional polymorphisms that will be used in future population genetic studies in free-range gorilla. Keywords: nuclear translocation, mitochondrial DNA, numt, great ape, primate, evolution, conservation genetics.

GENERAL INTRODUCTION

The story of mitochondrial colonization of the nuclear genome

The origin of the mitochondrion and genetic exchange with the nucleus. Once free-living prokaryotes, α-proteobacteria gave rise to mitochondria through endosymbiosis with eukaryotic cells (Gray & Doolittle 1982). Since then, an intensive communication process between this organelle and the nucleus has involved exchange of molecules including genetic material that has persisted up to these days (Timmis et al. 2004). Endosymbiosis resulted in loss of genes that were no longer needed such as those for synthesis of the bacterial cell wall (Adams & Palmer 2003). While some genes were also lost because of their functional redundancy, other genes were relocated to the nucleus where they are expressed to serve mitochondrial requirements or where they evolved new functions (Adams & Palmer 2003; Timmis et al. 2004). An exceptional case is the proton-translocating ATPase of the fungus *Neurospora crassa*, which is still expressed by both the nuclear and mitochondrial genomes (van den Boogaart et al. 1982). Also, recent transfer of functional mitochondrial genes to the nuclear genome has been shown in flowering plants (Boore 1999; Leister 2005; Adams et al. 2000). However, the translocation of functional genes seems to have ceased in animals due to changes in the mitochondrial genetic code relative to the universal code of the nucleus, leading to a loss of gene function once animal mitochondrial genes are relocated in a nuclear context (Gray 1999; Boore 1999; Adams et al. 2000). Therefore, the majority of functional nuclear genes in animals that originated through mitochondrial translocations may be traced to a period prior to the divergence of plants and animals.

With the exception of the Mitochondrial Control Region (MCR), intergenic regions and introns have been expelled from the mitochondrial genome of animal taxa. The number of genes has stabilized at only 12 to 13 protein coding genes, 22 tRNAs and 2 rRNAs (Boore 1999; de Grey 2005), revealing a steady evolutionary trend toward size reduction (Selosse et al. 2001; Bensasson et al. 2001). One hypothesis that explains contraction of the mitochondrial genome states that the high mutation rate in the animal mitochondrion, lack or limited recombination and the four-fold reduction in effective population size of the mitochondrion relative to the nuclear genome can enhance random fixation of deleterious mutations and magnify the effect of potential disadvantageous mutations (Blanchard & Lynch 2000; Selosse et al. 2001). This contrasts with nuclear genes where recombination and selection could purge deleterious mutations more effectively. A different trend is observed in plant mitochondria whose mutation rates are much lower than in the nucleus, recombination between heterotypic mitochondrial genomes is frequent and mitochondrial genomes remain relatively larger (Henze & Martin 2001). An alternative explanation for the small size of the animal mitochondrial genome, aside from the lack of any introns and intergenic regions, suggests that organelles with small genomes may replicate their genomic material faster and consume less energetic resources, leading to a more economical and selectively advantageous cell (Berg & Kurland 2000; Selosse et al. 2001).

The journey from the mitochondrion to the nucleus

Escape of DNA from the mitochondrion. The trend towards a reduced size of animal mitochondrial genomes in conjunction with the contrasting plasticity of nuclear genomes to accommodate DNA with no apparent functional or structural roles paved the road for the integration of fragments of mitochondrial DNA (mtDNA) into nuclear genomes, commonly known as numts or nuclear copies of mtDNA (Lopez et al. 1994). But the journey of a mtDNA fragment to the nucleus involves several steps starting with the release of mitochondrial fragments followed by their escape to the cytoplasm, importation into the nucleus and finally integration into the nuclear genome.

As part of the natural mitochondrial dynamics, the organelle suffers cycles of fission and fusion (Twig et al. 2008a; 2008b). After fission, fragments that have been depolarized due to defective functioning or programmed differentiation are targeted for degradation through a process of mitochondrial turnover called mitophagy (Abeliovich 2007; Kanki & Klionsky 2010). Increased mitophagy has been shown in yeast as a mechanism for removing damaged mitochondria under conditions of cell stress that may also increase the chance of mtDNA escape (Thorsness & Fox 1993; Shafer et al. 1999; Mijaljica et al. 2007). Mutations in a group of nuclear genes collectively called *yeast mitochondrial escape* or *YME* have been shown to affect mitochondrial morphology, impair vacuolar degradation of mitochondria and promote an elevated rate of mtDNA escape to the nucleus (Campbell & Thorsness 1998; Shafer et al. 1999; Priault 2005; Park et al. 2006; Abeliovich 2007). Increased mitochondrial escape to nucleus has also been shown in HeLa (cancer) cells and rat hepatoma cells (Corral et al. 1989; Shay &

Werbin 1992) and it has been shown to be correlated with an overproduction of mtDNA in brain tumors (Liang 1996). These observations suggest the disruption of physiological mechanisms of the cell that is possibly associated with incomplete degradation of mitochondrial debris and breaches in the vacuolar membrane. But the connections between the mentioned nuclear mutations, mitochondrial disfunction and transit of mtDNA to the nucleus remain to be uncovered (Thorsness & Weber 1996). However, the experimental data necessary for understanding the mechanisms and molecules involved in targeting specific mitochondria during the mitophagy of higher eukaryotes is still scarce and identification of genes orthologous to *YME* or other nuclear genetic variants accounting for the escape of mtDNA to the nucleus remain to be identified (Goldman et al. 2010; Kanki & Klionsky 2010).

It is well established that sperm mitochondria in most species with sexual reproduction are eliminated early during the embryogenesis. They are marked with ubiquitin and degraded by proteasomes and lysosomes (Sutovsky et al. 1999; Rawi et al. 2011; Sato & Sato 2011). But incomplete degradation of mtDNA caused by failure to efficiently mark paternal mitochondria or activate the process of autophagocytosis could be the first step in the process of release and transit of paternal mtDNA to the nucleus (Woischnik & Moraes 2002). In fact, occasional "leakage" of paternal mitochondria has been reported in a wide variety of taxa such as mussels (Zouros et al. 1992), flies (Kondo et al. 1990), birds (Kvist et al. 2003), and mammals (Gyllensten et al. 1991; Schwartz & Vissing 2002). Since numts have to be acquired within the germline in multicellular organisms in order to be inherited, sperm mitochondria could be a feasible source of numts (Willett-Brozick et al. 2001). How uptake of nucleic acids into the nucleus takes place is not clear. However, several scenarios could be hypothesized including DNA passage through transient membrane gaps or illicit importation, direct contact of mitochondrial and nuclear membranes and encapsulation of mitochondrial compartments inside the nucleus (Thorsness & Weber 1996; Shafer et al. 1999; Hazkani-Covo et al. 2010).

Nucleic acids are thought to migrate to the nucleus in the form of genomic DNA or cDNA (Nugent & Palmer 1991; Henze & Martin 2001). Protein-coding genes in plants often have introns that are absent in nuclear copies of these genes indicating that mitochondrial integration is mediated by cDNA derived from spliced mRNAs (Nugent & Palmer 1991; Henze & Martin 2001; Adams & Palmer 2003). However, mitochondrial introns in plants can be mobile and nuclear translocations may have occurred at a time when the genes did not possess introns (Malek et al. 1997; Henze & Martin 2001). In contrast, direct DNA transfer has been experimentally demonstrated in yeast and higher eukaryotes, as evidenced by the presence of numts extending across two or more genes or derived from non-coding mitochondrial regions, collectively suggesting that any portion of the mitochondrial genome can be transferred to nucleus (Thorsness & Fox 1990; Henze & Martin 2001; Woischnik & Moraes 2002).

Mitochondrial DNA becomes part of the nuclear genome. Once in the nucleus, the chromosomal integration of imported mtDNA is thought to take place through two possible mechanisms. Mitochondrial fragments are thought to opportunistically insert in double strand breaks (DSB) in the nuclear genome guided by microhomologies between fragments that are then ligated through a mechanism called Non-Homologous End-Joining Repair (NHEJR) intended to repair such DSBs (Blanchard & Schmidt 1996; Yu & Gabriel 1999; Willett-Brozick et al. 2001; Jackson 2002; Hazkani-Covo & Covo 2008). Secondly, *trans*-replication slippage may also mediate the integration of mtDNA in the nuclear genome but this seems to be an exception more than the predominant mechanism (Chen et al. 2005). In this case, the 3' end of a broken strand in the nuclear DNA (nDNA) dissociates from the template strand and misaligns with a mtDNA molecule via *trans*-sequence homology which is subsequently used as replication template. Then the primer strand dissociates from the mitochondrial template and re-anneals to the nuclear template strand via *trans*-sequence homology of short direct repeats.

In addition to these two mechanisms, an apparent close proximity of Transposable Elements (TEs) to numts supports the idea that TEs might also mediate numt insertion (Farrelly & Butow 1983; Ricchetti et al. 1999; 2004; Mishmar et al. 2004; Lascaro et al. 2008). In humans for instance, Mishmar et al. (2004) found that a particular family of TEs called Long Interspersed Elements (LINEs) integrate within 150bp of numts. However, these findings contrast with more recent studies that have revealed an apparent deficit of TEs within 200bp of numt loci (Gherman et al. 2007; Jensen-Seaman et al. 2009). Interestingly, LINE-1 (or L1), which are the most abundant retrotransposons in the human genome, have the ability to transduce genomic regions on the 3' flank thus allowing their duplication and insertion elsewhere in the genome (Moran et al. 1999; Pickeral et al. 2000; Goodier et al. 2000; Deininger et al. 2003; Xing et al. 2006). Numt duplication through L1 3' transduction is supported by the predominant non-contiguous genomic location of numt duplicates and partial evidence showing the physical association of LINEs and numts (Hazkani-Covo et al. 2003; Bensasson et al. 2003; Triant & DeWoody 2007). Although this mechanism of insertion implies an elevated proportion of numt duplications vs. independent mitochondrial integrations, as it has been reported in several studies (Lopez et al. 1994; Collura & Stewart 1995; Tourmen et al. 2002; Antunes & Ramos 2005; Pamilo et al. 2007; Behura et al.

2007; Triant & DeWoody 2007; Hazkani-Covo et al. 2003), overall evidence suggests that most numts are generated through direct mitochondrial integration (see Mishmar et al. 2004).

Bioinformatic and wet-bench experiments have shown that numts are scattered throughout the chromosomes in mammals (Gherman et al. 2007). Like LINEs, numts also insert preferentially in non-coding regions with GC-poor isochores (Mishmar et al. 2004), suggesting selective pressure against structural and functional disruption of active genes (Saccone et al. 2002; Lascaro et al. 2008; Mishmar et al. 2004). However, deleterious effect of numts can still be detected in a limited number of recent human numts inserted into functional parts of the genome. Examples of this are: 1) a 41bp mitochondrial fragment inserted at the breakpoint junction of a reciprocal constitutional translocation, segregating in a family with bipolar disorder (Willett-Brozick et al. 2001). 2) a 251-bp insertion causing a bleeding disorder (Borensztajn et al. 2002). 3) A *de novo* mitochondrial insertion of 72bp that causes a rare condition of developmental disorders called Pallister-Hall syndrome (Turner et al. 2003). 4) A 93bp insertion that results in mucolipidosis type IV, a disorder characterized by delayed psychomotor development and visual impairment (Goldin et al. 2004). 5) A 36bp insertion that causes a type of deaf-blindness called Usher syndrome (Ahmed et al. 2002; Chen et al. 2005). 6) At least three numt polymorphisms inserted in known genes (Ricchetti et al. 2004).

Numts are not equally abundant in all genomes

Since the initial recognition of nuclear mitochondrial sequences in the mouse genome over four decades ago (Du Buy & Riley 1967), numts have been found in abundance in a wide range of taxa (Adams et al. 2000; Bensasson et al. 2001; Richly & Leister 2004; Triant & DeWoody 2007; Sacerdot et al. 2008; Nergadze et al. 2010) (see Figure i). With the advent of technological advances in massive genomic sequencing and advances in bioinformatic tools, several research groups have estimated the prevalence of numts in individual genomes and made comparisons across taxa (Richly & Leister 2004; Triant & DeWoody 2007; Hazkani-Covo et al. 2010; Lang et al. 2011). Recent comparisons of numt content across all major eukaryotic taxa revealed a positive correlation between genome size and numt content (Hazkani-Covo et al. 2010). Since non-coding DNA accounts for major differences in the genome size in eukaryotes (Kidwell 2002), one would expect to find not only more non-coding DNA but also more numts in larger genomes (Bensasson et al. 2001; Hazkani-Covo et al. 2010). Although evidence from humans has shown that numts are occasionally found in coding or regulatory regions, they randomly insert across the genome (Gherman et al 2007), suggesting that larger genomes would confer more opportunities for numts to insert.

Surprisingly, a great deal of variation in numt content has been observed between related taxa of several animal groups, including mosquitoes (*Anopheles gambiae* and *Aedes aegypti*) (Pamilo et al. 2007; Black IV & Bernhardt 2009), carnivores (dogs and cats) (Triant & DeWoody 2007; Antunes 2007); rodents (mice and rats) (Triant & DeWoody 2007) and Old World primates (macaques and chimpanzees) (Triant & DeWoody 2007). Such disparities may stem from different demographic histories (Gherman et al. 2007), species-specific mechanisms controlling mitochondrial escape or differences in the stability of the nuclear genome (Richly & Leister 2004; Leister 2005). Interspecific differences in numt content could also be an artifact caused by different genomic search strategies. In general, estimates of numt content will increase with more relaxed parameter settings since ancient, highly divergent or small insertions may be particularly hard to detect. However, this strategy easily leads to spurious associations and thus false numt hits. Even independent numt searches conducted at different times in the same species may yield striking differences in numt counts (see Figure i), which again may be caused by different parameter settings or the completeness of the genomic database under study. For instance, an increase in human numt content from 279Kbp in 2004 (Richly &Leister) to 406Kbp in 2007 (Hazkani-Covo & Graur) indicates either further numt discoveries as sequencing of the reference genome progressed or relaxation of settings in the numt search. But a reduction to 264Kbp in 2010 (Hazkani-Covo et al.) is mostly explained by the use of more stringent parameters in the numt search of this last study. Due to the limitations of bioinformatic tools, further validation of candidate numts should be done. For instance, match of a nuclear sequence with two or more mitochondrial regions is an indication that at least one of the matches could be spurious. Whenever possible, regions containing low-score hits such as small or highly divergent sequences should be aligned with related genomes to verify that putative numts match these insertions (Zischler et al. 1995b; 1998; Hazkani-Covo & Graur 2007).

Pros and cons of numts in evolutionary population studies

Contamination of mitochondrial databases. Recent or highly conserved numt sequences usually exhibit great similarity with modern mtDNA and pose an imminent risk of inadvertent amplification with mitochondrial primers and contamination of mitochondrial databases (Zhang & Hewitt 1996a; Jensen-Seaman et al. 2004; Richly & Leister 2004; Pamilo et al. 2007; Triant & DeWoody 2007). Since mitochondrial sequences are heavily used in population genetics and systematics, misidentification of numts as mitochondrial sequences has led to overestimation of mitochondrial diversity (Garner & Ryder 1996; Song et al. 2008; Moulton et al. 2010; Bertheau et al. 2011), incorrect phylogenetic analyses (Hedges & Schweitzer 1995; van der Kuyl 1995) and misdiagnosis of mitochondrial genetic disease (Hirano et al. 1997; Wallace et al. 1997).

The risk of amplifying numts using mitochondrial primers is worsened by multiple factors starting with the primers themselves. The potential of generic, also called universal, primers to amplify conserved mitochondrial regions in non-target species is undeniable (Kocher et al. 1989; Naidu et al. 2012), but they should be used with caution since they also tend to anneal with nuclear pseudogenes. Since nuclear copies usually mutate at slower rates than their mitochondrial paralogs and represent ancient mitochondrial lineages they are also known as mitochondrial “molecular fossils” (Perna & Kocher 1996). As a consequence, numts may compete or even impede amplification of mitochondrial templates during the PCR (Arctander 1995; DeWoody et al. 1999; Thalmann et al. 2004; Grosso et al. 2006; Podnar et al. 2007). Another complication comes from the type and storage of the biological tissue used as source of DNA. Genetic studies of wildlife frequently make use of hair, feces and museum specimens but it has been observed that depending on the tissue, environmental/storage conditions or DNA extraction protocol, stability of mitochondrial and nuclear DNA may decay at different rates (Wallace et al. 1997; Berger et al. 2001; Castella et al. 2006). It has been determined that mtDNA generally disintegrates faster in feces or poorly stored hair and soft tissues (Berger et al. 2001; Roon et al. 2003; Foran 2006; Soto-Calderón et al. 2009), hence increasing the nuclear-to-mitochondrial ratio and the chance of amplifying nuclear templates. Greenwood and Pääbo (1999) have shown in elephants for instance that primers intended to amplify MCR do so from blood DNA but preferentially amplify nuclear copies from hair DNA.

The potential use of numts in systematics and population genetics. Numts also exhibit several properties that can be exploited for evolutionary studies. As numts represent “fossilized” copies of ancestral mtDNA haplotypes, they may serve as outgroups in phylogenetic analysis where other suitable outgroups are unavailable (Bensasson et al. 2001; Zischler et al. 1995b; Hay et al. 2004). Also, numt loci may also be useful as phylogenetic or population genetic markers (Zischler et al. 1998). Similar to TEs, numts occasionally exhibit insertional polymorphisms with great potential as binary markers in systematics (Ray et al. 2006; Herke et al. 2007), markers for forensic identification of species (Walker et al. 2003) or as population genetic markers (Perna et al. 1992; Watkins et al. 2003; Schmitz et al. 2005).

Contrary to other codominant markers such as microsatellites and RFLPs, TEs and numts are considered free of homoplasy as once inserted they are rarely excised from the nuclear genome. Therefore identical insertions may additionally be considered identical by descent (Batzer & Deininger 2002). Since the ancestral allele is considered to be the absence and the derived allele the presence, these binary markers could also be used to assess not only intensity but also directionality of migrational patterns in natural populations (Thomas et al. 1996; Batzer & Deininger 2002). Identification of geographic structure in variable human numts has paved the road for their use in the study of human evolution (Giampieri et al. 2004; Yuan et al. 1999; Lang et al. 2011). This is illustrated by frequency gradient of a 540bp human numt in populations around the world (Thomas et al. 1996), which reveals a pattern that is consistent with the hypothesis of African origin of human populations and their subsequent dispersal to Eurasia and the Americas.

A mitochondrial fragment inserted in a specific genomic location represents a snapshot of a past insertional event that involved a particular mitochondrial haplotype and a specific genotype of the flanking region, both coexisting in a geographical region at the time of the insertion. This association may give clues about the ancestral co-distribution of both mitochondrial and nuclear variants, geographic origin of numts and the historical structure in ancient populations (Hazkani-Covo 2010). Apart from humans and a few other taxa (Thomas et al. 1996; Nergadze et al. 2010; Miraldo et al. 2012), the utility of numts in evolutionary studies has been underexplored but future advances in genome sequencing as well as development and implementation of molecular tools will certainly allow a wider use of these genetic elements.

Isolation and avoidance of numts. Given the widespread distribution of numts, their similarity to authentic mitochondrial sequences and their inadvertent amplification in studies of mtDNA, it is necessary to implement quality control methodologies aimed at systematically avoiding amplification of such copies or contrarily, attain their effective isolation and characterization.

Overlapping peaks in sequencing profiles of mitochondrial genes frequently indicate co-amplification with nuclear pseudogenes and/or heteroplasmy (i.e. multiple populations of mtDNA) (Chinnery et al. 2000; Thalmann et al. 2004; McLeod & White 2010). When translocated to the nucleus, numts escape from mitochondrial selective pressures (Smith et al. 1992; Bensasson et al. 2001; Schmitz et al. 2005) and accumulate mutations that are rarely observed in functional mtDNA. In the case of protein-coding mitochondrial genes, nuclear copies suffer missense and non-sense mutations as well as indels. Also, co-occurrence of two or more different mitochondrial genomes, a phenomenon called heteroplasmy, can be caused by mutation in the female germline, paternal mitochondrial "leakage" or direct maternal inheritance (Gyllensten et al. 1991; Jenuth et al. 1997; Chinnery et al. 2000; Kvist et al. 2003; Calloway et al. 2000). Isolation of several mitochondrial-like sequences from the same individual is also evidence of numts or heteroplasmic mtDNA (Garner & Ryder 1996; Mundy et al. 2000). As numts diverge from mitochondrial sequences both in the pattern and rate of nucleotide substitution (Arctander 1995; Lopez et al. 1997; Schmitz et al. 2005), unusually long branches in mitochondrial phylogenetic analysis may actually correspond to unexpected co-amplification of numts sequences (Zischler et al. 1998; Jensen-Seaman et al. 2004). Occasionally, additional PCR recombinants of mitochondrial and nuclear templates may also be co-amplified with native templates and increase the diversity of products (Saiki et al. 1988; Pääbo et al. 1990; Anthony et al. 2007a). This happens when the polymerase switches between different templates, and may be exacerbated by excessively long PCR programs and primer depletion (Judo et al. 1998; Thalmann et al. 2004). Amplification of these chimerical products is stochastic so that the same sequence generally fails to amplify more than once. Multiple amplification trials under varying cycling conditions can then be tried to identify or even isolate these sequences. Also, recombination detection methods have proved effective in detecting candidate recombinants in a pool of mitochondrial sequences (Anthony et al. 2007a).

Several laboratory methods are currently available to avoid numts and distinguish them from mitochondrial sequences. Mitochondrial enriched DNA samples may be obtained through CsCl gradients or commercial DNA isolation kits (Zhang & Hewitt 1996b; Ibarguchi et al. 2006 for details), but this should be complemented with further methods to attain specific mtDNA amplification. Since the size of most numts is below 500bp one can amplify (Richly & Leister 2004; Pamilo et al. 2007; Gherman et al. 2007), even from extracts of total DNA, several thousand base pairs of mtDNA through Long-Range PCR (LR-PCR) and then use internal primers for nested PCR or direct sequencing (Thalmann et al. 2004; Calvignac et al. 2011). Because of the greater ratio of mitochondrial to nuclear DNA in fresh samples, serial dilutions of

DNA extracts are expected to dilute out nDNA and favor amplification of mitochondrial templates (Ibarguchi et al. 2006). Finally, RT-PCR may also be considered for specific amplification of mitochondrial cDNA given the presumed lack of transcriptional activity of animal mitochondrial pseudogenes (Sunnucks & Hales 1996).

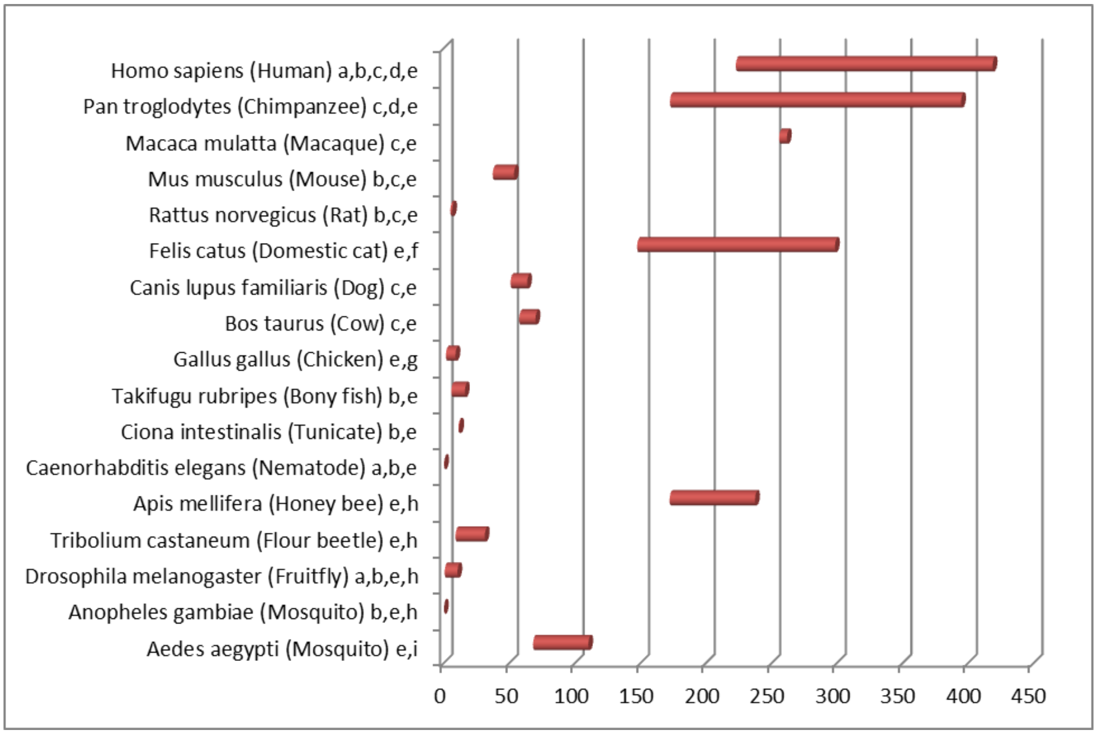
Alternatively, isolating numts is desired when the interest is for example estimating numt prevalence, numt mechanisms of insertion, sequence divergence from mtDNA or testing the distribution of polymorphic insertions in a population. BLAST searches are probably the most straightforward way to identify and map numts. But comprehensive databases are only available for a limited number of taxa and they are solely based on one individual so complementary bench experiments are needed to increase the chance of capturing numts with insertional polymorphisms absent in reference genomic databases. One alternative is cloning and sequencing PCR products of mtDNA which allows the identification of potential numts and assessment of primer specificity (Mundy et al. 2000; Vallinoto et al. 2000; Thalmann et al. 2004; Moulton et al. 2010). In fact, amplification with generic or degenerate primers has been used to deliberately promote co-amplification of mitochondrial and nuclear products and can be an effective way to recover numts (Sunnucks & Hales 1996; Bensasson et al. 2000; Mundy et al. 2000; Williams & Knowlton 2001; Thalmann et al. 2004). Also, Fluorescent *In Situ* Hybridization (FISH) has been used as an approach to detect and visualize the distribution and abundance of numts in chromosomes (Gherman et al. 2007). But the limitation of all these approaches however is the inability to map the genomic location of amplified sequences and therefore discriminate allelic variants of one locus from amplification of independent loci. Although a wide suite of methods have been successfully used in the past for the identification of sequences flanking target genetic elements, they have been underexploited in the case of numts. These include chromosome-walking methods (Ochman et al. 1988; Jones & Winistorfer 1992; Zischler et al. 1995b; 1998; Yuanxin et al. 2003; Tan et al. 2005; Ray et al. 2005; Ren et al. 2005) and methods based on next-generation sequencing (Mardis 2008; Hudson 2008; Mason et al. 2011).

Numts in Old World primates

Sequenced genomes of Old World primates such as macaques and great apes have a high abundance of numts (chimpanzees, humans, orangutans, gibbons and gorillas) (Vartanian

& Wain-Hobson 2002; Jensen-Seaman et al. 2004; Anthony et al. 2007a; Chung & Steiper 2008; Hazkani-Covo 2009; see also Figure i). Multiple factors make this group key in the study of numts. First, the availability of genomic databases of great apes in particular has facilitated the identification of numts, which is of great utility in the study of mitochondrial-nuclear communication, modes of mitochondrial integration in the nucleus and numt evolution. Also, numt prevalence has been shown to be elevated in the genome of multiple primate taxa to the point that profuse contamination of mitochondrial databases, as is the case of gorillas, has led to questioning the reliability of these databases. This leads to several questions of interest and a few could be stated the following way: why are numts so common in primates? What are the causes of differences between closely related species? How common are numts in unsequenced or partially sequenced genomes? How fast do numts and mitochondrial paralogs diverge? What is the actual risk of numt amplification with reported primers?

Figure i. Numt content (Kbp) in animal genomes. Bars indicate the range in the estimated numt content from independent studies (a-k). These ranges are wider in genomes that seem to have greater numt content (E.g. human, chimpanzee and domestic cat).



(a) Bensasson et al. 2001; (b) Richly & Leister 2004; (c) Triant & DeWoody 2007; (d) Hazkani-Covo 2007; (e) Hazkani-Covo 2010; (f) Antunes 2007; (g) Pereira and Baker 2004; (h) Pamilo et al. 2007; (i) Black IV & Bernhardt 2009; (j) Sacerdot et al. 2008; (k) Lenglez et al. 2010

Estimating the rate of numt insertion. The number of numts is a combination of *de novo* integration or post-integration duplication of existing numt loci. Several studies argue that the rate of numt insertion has been constant during the diversification of Old World primates including great apes (Mourier et al. 2001; Hazkani-Covo et al. 2003), whereas others indicate this rate has varied reaching a peak early during diversification of Old World and New World primates (Bensasson et al. 2003; Gherman et al. 2007).

But several factors that deserve to be mentioned here may affect the estimation of insertion rates. First, estimation of ancient insertion rates is challenging due to the loss in identity of old numts relative to contemporary mitochondrial sequences and the cumulative effect of genomic reorganizations and deletions that can potentially erode any trace of the original insertion. In humans for instance, an observed deficit in the number of MCR numts compared to other mitochondrial regions could stem from detection bias arising from the high mutation rate of MCR and its rapid loss in sequence identity with nuclear copies (Saccone et al. 1991; Sbisà et al. 1997; Mourier et al. 2001). Since MCR is the only mitochondrial region that does not transcribe (Sbisà et al. 1997; Fernandez-Silva et al. 2003), a correlation between abundance of mitochondrial transcripts and the number of nuclear copies could in principle explain the deficit of MCR numts. Although tempting, this hypothesis is unlikely as such correlation does not seem to exist in humans (Woischnik & Moraes 2002). Alternatively, the elevated mutation rate of the mitochondrial MCR and its rapid loss of sequence identity could explain the MCR numt observed deficit (Mourier et al. 2001; Woischnik & Moraes 2002). This might also mean that the true impact of mitochondrial transfers in shaping the architecture of the nuclear genome could be easily underestimated; an effect that has not been fully evaluated.

In contrast to old numts, the rate of recent numt insertion may be directly estimated from their prevalence in each species. However, this seems to vary even between closely related taxa suggesting an effect of factors intrinsically associated with physiological mechanisms or the demographic history of a given species (Hazkani-Covo & Graur 2007; Hazkani-Covo 2009). That is the case of a historical bottleneck that is deemed to have caused a reduced genetic diversity in human populations as compared with chimpanzees and other great apes with larger historical effective population sizes (Kaessmann et al. 2001; Gherman et al. 2007; McEvoy et al. 2011). Also, differences in mechanisms of mitochondrial integration or genome reorganization might hypothetically cause the apparent differences in numt prevalence across species but further evidence of the role of these factors remain to be gathered.

Numts may be the product of independent integration of mitochondrial fragments or duplication of established numts (Collura & Stewart 1995; Lopez et al. 1996; Stupar et al. 2001; Bensasson et al. 2003). Identification of duplication events should be straightforward when the identity between two numts can be traced to the flanking regions as in the case of duplicated chromosomal fragments (Lopez et al. 1994; Bensasson et al. 2003; Hazkani-Covo & Graur 2007). But partial duplication of an internal numt fragment may be particularly hard to distinguish from independent integrations of similar or identical mitochondrial haplotypes (Bensasson et al. 2003). That a long and a short numt appear as sister taxa in a phylogenetic analysis may indicate that the long numt gave rise to the short numt (Hazkani-Covo et al. 2003), but the same pattern may also be obtained from independent integrations of related mitochondrial haplotypes (Bensasson et al. 2003).

Phylogenetic methods have also been widely used to place numts in a reference mitochondrial phylogeny and in this way infer times and rates of insertion (Collura & Stewart 1995; Bensasson et al. 2003; Hazkani-Covo et al. 2003; Gherman et al. 2007). But this practice is problematic since the location of a numt duplication in the tree will be influenced by the age of the original integration rather than the date of the duplication event. Rates and patterns of nucleotide substitution are substantially different in mitochondrial and nuclear sequences (Arctander 1995; Lopez et al. 1997; Schmitz et al. 2005), and their incorporation in the same phylogenetic analysis may be therefore misleading and result in incorrect placements, assigning numts to taxa where it is not present and creating unexpectedly long branches connecting to numts in the tree (Graur & Li 2000; Schmitz et al. 2002; 2005; Podnar et al. 2007). Additionally, numts generally contain insufficient phylogenetic information to accurately place their time of insertion due to their small size (usually <500bp) and slow substitution rate (Jensen-Seaman et al. 2009). Therefore, caution should be used when origin of a numt is assigned through phylogenetic inference and complementary methods such as direct inspection of presence/absence patterns in target taxa should be used whenever possible (Ray et al. 2005; 2006).

Evolution of homologous sequences in two different cell compartments. The nuclear and mitochondrial genomes differ in many aspects including topological organization, mode of replication, patterns of selection and the types of repair mechanisms among other aspects (Fernández-Silva et al. 2003; Meiklejohn et al. 2007). The transplantation of mitochondrial sequences to the nucleus sets up the conditions for a natural experiment for assessing the effect of two different intracellular environments on the evolution of homologous sequences. The mitochondrial genome of Old World primates for instance, has an upward bias in GC content that exceeds the levels observed in other mammals and is apparently led by lineage-specific mutational pressure (Schmitz et al. 2002). Also, the elevated substitution rate and strong transition-biased nucleotide substitution pattern are common trends of the vertebrate mitochondrial genome that combined lead to saturation in the number of transitions and underestimation of transition/transversion (Ts/Tv) ratios (Arctander 1995; Lopez et al. 1997; Purvis et al. 1997; Yang & Yoder 1999). In contrast, nuclear copies remain relatively conserved (Brown et al. 1982; Graur & Li 2000; Haag-Liautard et al. 2008), and their escape from the mutational bias may make them behave as snapshots of the mitochondrial sequence that reflect the GC content at the time of translocation (Perna & Kocher 1996; Bensasson et al. 2001; Zischler et al. 1995b).

The case of numts in gorillas. Since the formation of numts seems to be an ongoing process in Old World primate genomes, the resemblance of recent numts to contemporary mitochondrial genomes poses a potential risk of contamination of mitochondrial databases (Song et al. 2008; Calvignac et al. 2011). But nowhere is this problem worse than in gorillas as it turns out that numerous sequences originally reported as MCR are actually nuclear translocations (Jensen-Seaman et al. 2004; Thalmann et al. 2004; 2005; Anthony et al. 2007a). This problem is aggravated by the apparently high incidence of *in vitro* recombinants of mitochondrial and nuclear templates that have also been misdiagnosed in previous studies (Anthony et al. 2007a). For instance, although the mitochondrial sequences corresponding to the first hyper-variable domain (HVI) of the western gorilla Rok and the lowland eastern gorilla Muk were recovered through LR-PCR, an additional number of mitochondrial-like sequences were amplified from the same individuals using standard PCR methods (Thalmann et al. 2004). Phylogenetic analyses of HVI sequences from gorillas across their range have recovered three different numt clusters (I – III) that are interspersed with four mitochondrial haplogroups A - D (Clifford et al. 2004; Anthony et al. 2007a; 2007b). Interestingly, the nuclear and mitochondrial copies of HVI exhibit high similarity and both bear a poly-C domain that is unique to gorillas emphasizing a burst in the origin of nuclear copies after the divergence of this taxon. All gorilla numt sequences obtained so far have been recovered through non-specific amplification with primers originally designed to amplify mtDNA, thus limiting the ability to amplify particular loci, assess numt diversity and reconstruct the steps of mitochondrial integration (Garner & Ryder

1996; LaCoste et al. 2001; Jensen-Seaman et al. 2004; Thalmann et al. 2004; 2005). Therefore, future advances in the gorilla genome project and numt mapping will allow a detailed characterization of these genetic elements (Zischler et al. 1998; Scally et al. 2012).

Introduction to the following chapters

In the following chapters I address several aspects of the evolutionary dynamics of nuclear translocations of mtDNA in great apes. As presented above, these fragments are widely distributed in great apes and pose an imminent contamination threat for mitochondrial databases (Richly & Leister 2004; Jensen-Seaman et al. 2004; Triant & DeWoody 2007; Anthony et al. 2007a). Preliminary studies in gorillas suggest that the MCR translocation rate is much higher in this species than either chimpanzees or humans (Thalmann et al. 2005). However, a systematic inventory and rigorous comparative analysis across these closely related taxa is presently lacking. Since, MCR has been extensively used as molecular marker in population genetics, identification and characterization of MCR numts is essential in designing quality control measurements that prevent contamination of mitochondrial databases. The main goal of this thesis is therefore to compare the evolutionary dynamics of nuclear copies of the MCR in great apes (chimpanzees, humans, gorillas and orangutans).

Chapter 1. Factors affecting the relative abundance of nuclear copies of the mitochondrial control region (numts) in hominoids. Although genomic sequencing and experimental evidence have shown an elevated prevalence of numts representing all portions of the mitochondrial genome in great apes, the MCR seems to be underrepresented in the nuclear genome of humans relative to other mitochondrial regions (Mourier et al. 2001). Whereas this observation may be the consequence of an actual deficit in the number of translocations of this region, the most likely explanation of this apparent numt deficit is rapid loss of identity between mitochondrial and nuclear copies that would be caused by the elevated rate of evolution of this mitochondrial region. In this chapter I address the question of whether the apparent deficit in MCR numts observed in humans is a conserved pattern in other great apes. Since the erosion of sequence identity could account for the apparent deficit of numts from MCR, the same bias might also be evident for numts from more highly variable sub-domains within MCR. To answer these questions, BLAST was used to identify MCR numts in the reference sequenced genomes of humans, chimpanzees and orangutans. The insertion point of each numt was then inferred in the reference hominoid phylogeny based on the presence/absence pattern in each taxon and this information was used to estimate the rate of MCR numt insertion in each lineage. Lastly, the numt prevalence across the four MCR sub-domains (HV1, CCD, HV2 and MCR_F) was assessed to test the hypothesis that an apparent deficit in MCR numts is an artifact of rapid loss of sequence identity. If this is the case, the most variable sub-domains should exhibit the smallest number of detected numts.

Chapter 2. Nucleotide composition, sequence evolution and mechanisms of insertion of nuclear copies of mitochondrial DNA in great apes (Hominoidea). Once mtDNA fragments colonize the nuclear genome, they experience an environment that is substantially different from the conditions of native mitochondrial sequences, providing an unparalleled opportunity to compare the evolution of homologous sequences in mitochondrial and nuclear contexts. There are many structural and selective differences between these two genomes including selective pressures and different rates of nucleotide substitution (Fernández-Silva et al. 2003; Meiklejohn et al. 2007). As the effect of the genomic context is expected to be cumulative over time, comparisons of homologous sequences in these two genomes require not only the identification of a target numt population but previous knowledge of their age. Previous studies have only made use of a limited number of loci to address the effect of the two genomic contexts on the structure and evolution of nuclear and mitochondrial paralogous sequences. In this chapter, I build on a dataset of 83 numts inserted at different times in great apes since their divergence from the macaque lineage. The insertion point of each numt in a reference phylogeny was inferred from their presence/absence patterns in all major great ape taxa. Differences in GC content and the observed ratio of transition and transversions (T_s/T_v) for each numt and its mitochondrial copy were then examined. Since TEs may potentially elicit duplication of flanking regions, concordance in the insertion time of numts and neighboring TEs was also studied as a means of gathering indirect evidence of the potential role of TEs in numt duplication. In the previous chapter, I assessed the hypothesis that unusually high rates of sequence evolution in MCR account for the apparent deficit of numts in this region in great apes. Continuing with this idea, I evaluate in this chapter the hypothesis that more conserved mitochondrial genes should exhibit a relatively larger number of nuclear copies. To test this idea, complete mitochondrial genomes and number of numts derived from mitochondrial genes were compiled from previous publications.

Chapter 3. Isolation of novel nuclear insertions of mitochondrial DNA (numts) in gorillas and their potential as population genetic markers. Accidental amplification of HVI numts and *in vitro* recombinants has been so common in gorillas that validity of mitochondrial databases has been questioned (Jensen-Seaman et al. 2004; Thalmann et al. 2004; 2005). Three groups of gorilla numts have been described in both eastern and western gorillas suggesting a wide distribution of multiple numts. As no numts have been directly mapped, amplification of specific numts remains incidental and this impedes further analysis of numt diversity and implementation of measures to prevent numt contamination. The main purpose of this chapter is to isolate gorilla numts and determine whether the nuclear origin of previously inferred numt sequences can be confirmed in this way. The genomic location of each numt was characterized through three complementary methodologies: 1) Numt BLAST searches of the draft of the reference gorilla genome using the whole mitochondrial genome as query sequence; 2) Screening of a commercial genomic library of gorilla contained in Bacterial Artificial Chromosomes (BACs) and; 3) Anchored PCR from a sample of five unrelated gorillas enriched for nDNA. In addition to numt isolation, specific primers were designed to determine the polymorphic status of each numt in a sample of western lowland gorillas captive in US zoos and explore their potential utility as nuclear molecular markers for future population genetic studies.

CHAPTER 1

FACTORS AFFECTING THE RELATIVE ABUNDANCE OF NUCLEAR COPIES OF THE MITOCHONDRIAL CONTROL REGION (NUMTS) IN HOMINOIDS.

Abstract

Although nuclear copies of mitochondrial DNA (numts) can originate from any portion of the mitochondrial genome, evidence from humans suggests that nuclear insertions of the mitochondrial control region (MCR) are less abundant than translocations from other mitochondrial regions. This apparent deficit might arise from the erosion of sequence identity in numts originating from rapidly evolving sequences such as the MCR. The same bias may also be evident for numts from more highly variable sub-domains of the MCR. However, the extent to which sequence properties of different portions of the mtDNA impact estimates of numt abundance has not been rigorously evaluated. To address this question, we first conducted an exhaustive BLAST search of MCR numts in the three well-studied hominoid genomes (human, chimpanzee, and orangutan) and assessed numt prevalence across the four MCR sub-domains. The date of numt insertion in the hominoid phylogeny was then assessed by BLAT or cross-species PCR of other hominoid genomes. Results indicate a marked deficit of numts from the second hyper-variable region and subdomain proximal to the tRNA-Phenylalanine in all three species. Both MCR subdomains exhibited the highest proportion of variable sites and lowest average number of detected numts/site. Variation in MCR insertion rate between lineages was observed with a pronounced burst in recent insertions within the chimpanzee and the orangutan. Lastly, the most variable subdomains are under-represented in ancient numts (older than 25 Mega-annum; Ma). Consequently, most species-specific numts closely resemble their mitochondrial counterparts, further underlining the risk of their inadvertent incorporation into mitochondrial datasets of primates.

Introduction

Fragments of mitochondrial DNA (mtDNA) translocated into the nucleus (numts) are present in a wide range of eukaryotes (Du Buy & Riley 1967; Corral et al. 1989; Bensasson et al. 2001; Hazkani-Covo et al. 2010). Once integrated into the nucleus, numts escape from mitochondrial selective constraints and are thought to mutate at rates that resemble other nuclear loci, which are around one order of magnitude slower than the mitochondrial average (Brown et al. 1982; Haag-Liautard et al. 2008). For this reason, numts are usually considered “fossilized” copies of ancient mitochondrial lineages (Perna & Kocher 1996; Bensasson et al. 2001; Zischler et al. 1995b), whose inadvertent amplification can potentially contaminate mitochondrial databases (Greenwood & Pääbo 1999; Jensen-Seaman et al. 2004; Anthony et al. 2007a). This problem is particularly acute for the mitochondrial control region (MCR) given its widespread use as a population genetic marker in many vertebrate taxa including great apes (Sbisà et al. 1997; Jensen-Seaman and Kidd 2001; Arora et al. 2010). However, the prevalence of MCR insertions in many of these taxa is poorly understood yet could have important implications for the use and interpretation of population genetic datasets.

A numt search in an early draft of the human genome showed an apparent deficit in the number of MCR numts compared to other mitochondrial regions (Mourier et al. 2001). Two possible explanations have been proposed to explain this observation. One states that if numts are predominately derived from RNA transcripts then untranscribed portions of the mitochondrial genome, such as the MCR, will be under-represented in the nuclear genome. Although such a mechanism of genetic transfer to the nucleus has been previously shown in plants (Nugent & Palmer 1991; Henze & Martin 2001), it remains to be shown that this is also the case in animals (Lopez et al. 1994; Henze & Martin 2001; Mourier et al. 2001). Alternatively, the deficit of MCR numts might be due to a detection bias arising from the high mutation rate of MCR and hence rapid loss in sequence identity relative to other portions of the mitochondrial genome (Saccone et al. 1991; Sbisà et al. 1997).

The MCR is the only non-coding region in the mitochondrial genome and because of this might be more tolerant of indel events and nucleotide substitutions (Sbisà et al. 1997). Additionally, the MCR has a high prevalence of nucleotide repeats (i.e. low DNA complexity), which are known to have an elevated mutation rate (Bodenteich et al. 1992; Sbisà et al. 1997; Zardoya & Meyer 1998). Over time, these properties of the MCR domain might erode the mitochondrial sequence identity of the nuclear copies and thus explain the apparent numt deficit. Similarly, within the MCR, nucleotide variability and levels of DNA complexity are likely to differ among the four MCR sub-domains, potentially leading to differences in their apparent abundance in the nuclear genome. Specifically the vertebrate MCR domain is comprised of two hyper-variable regions (HV1 and HV2), a conserved central domain (CCD) and a terminal portion adjacent to the $tRNA_F$ (MCR_F). In the mitochondrial genome of mammals, the sub-domains HV2 and MCR_F exhibit considerable variation in not only nucleotide sequence composition and length but also in the proportion of repeat motifs (Sbisà et al 1997). If those mitochondrial regions with greater variation in great apes also exhibit a greater deficit in the number of numts relative to less variable regions, then the disparity in the abundance of numts from different mitochondrial regions might be explained by the greater difference in sequence identity between these mitochondrial domains and their nuclear copies.

The rate of transfer of mtDNA to the nuclear genome is also thought to have varied during primate evolution (Bensasson et al. 2003; Gherman et al. 2007). The fact that some numts exhibit insertional polymorphism also suggests that nuclear integration is an ongoing process in many species (Thomas et al. 1996; Ricchetti et al 2004; Anthony et al. 2007a). A critical step in gauging insertion rates is the reliable inference of numt age. Although phylogenetic methods have been traditionally used to date numts and estimate insertion rates in great apes (Bensasson et al. 2003; Hazkani-Covo et al. 2003), such approach can be misleading since numts are small (<500bp) and usually contain insufficient phylogenetic information to accurately place their time of insertion (Jensen-Seaman et al. 2009). Furthermore, estimating the time of insertion of numt loci is problematic when both mitochondrial and nuclear loci are combined into the same phylogeny due to striking differences in patterns and rates of nucleotide substitution between the nuclear and mitochondrial genomes (Graur & Li 2000; Schmitz et al. 2002; 2005). Alternatively, the approximate time of insertion of candidate loci in a reference phylogeny can be estimated by either conducting BLAST surveys of taxa which have whole genomic sequences available or via cross-species PCR amplification of candidate loci from taxa that presently lack a comprehensive genomic database (Zischler et al. 1998; Jensen-Seaman et al. 2009).

Given our present lack of understanding of the molecular evolutionary dynamics of great ape MCR numts and the importance of these genetic elements in mitochondrial genetic studies, the present study set out to conduct a rigorous inventory of MCR numts from reference genomic databases of human, chimpanzee and orangutan. These data were then used to compare the prevalence of numts from different sub-domains within the MCR in order to test the hypothesis that heterogeneity in the number of numts across MCR sub-domains could be explained by differential loss of sequence identity between mitochondrial and nuclear copies. If true, the preponderance of numts from each sub-domain would be negatively related to the proportion of variable sites and positively related to DNA complexity. The MCR numt loci obtained from this study were then used as query sequences in genomic surveys of other great ape taxa (gorilla and gibbon) in order to estimate their approximate time of insertion and test the hypothesis that the rate of numt insertion has been constant throughout the evolution of great apes (Hominoidea). These data were also used to determine whether more variable sub-domains are proportionally under-represented in more ancient numts. This research will ultimately contribute to a better understanding of the factors determining the apparent abundance and distribution of mitochondrial fragments in the nuclear genome of great apes and may have important implications for population genetic analyses of mtDNA where detection and elimination of numt contaminants is an issue.

Materials and Methods

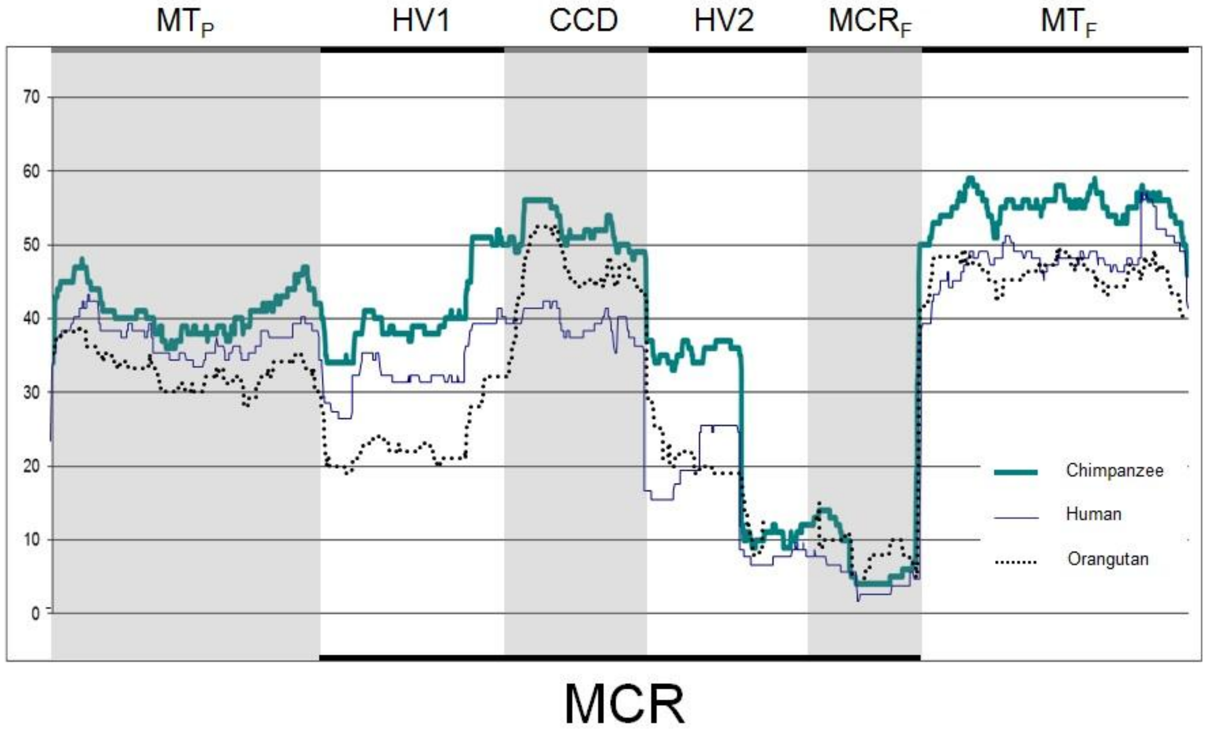
Relative abundance of MCR numts in the genome of humans, chimpanzees and orangutan. The BLASTn algorithm (Altschul et al. 1990) was used to carry out an exhaustive search for MCR numts in reference genome databases from human (build 36.3; 2006), chimpanzee (build 2.1; 2006) and orangutan (P_pygmaeus2.0.2;2007) assemblies. The MCR query sequence was taken from reference mitochondrial genomes of the corresponding species (NC001807.4 for human, NC001643.1 for chimpanzee and D38115.1 orangutan) and contains four sub-domains: the two hyper-variable regions (HV1 and HV2), the conserved central domain (CCD) and the sub-domain proximal to tRNA$_F$ (MCR$_F$). The query sequence employed in the present study also contained the two 500bp flanking regions, defined here as MT$_P$ and MT$_F$, where the former comprises tRNA$_P$, tRNA$_T$ and 32% of the *CYTB* gene, and the latter comprises tRNA$_F$ and 45% of 12S rRNA. A fragment of 81bp was found to be missing from the HV1 region of the mitochondrial reference sequence for the orangutan and so was replaced by another HV1 sequence reported in the same species (AJ586559.1). The filters and mask options of BLAST searches were clicked off; search parameters were relaxed to a word size of 7; match/mismatch scores of 1/-1 were adopted and gap creation and extension penalties of 3 and 1 were applied, respectively. Only hits of either i) at least 100bp in length and 60% identity or ii) a size of between 50 and 99bp with identity greater than 70% were considered. As preliminary analyses indicated that expect-values for discontiguous numt hits did not exceed 0.39, this value was used as an upper limit above which hits were rejected. This search strategy allowed us to recover all previously reported numts of at least 50bp in size as well as a number of unreported MCR numts.

Abundance of MCR numts across the different sub-domains. The mitochondrial sequences of the four major taxa in the Hominidae, i.e. human, chimpanzee, orangutan (D38115.1-AJ586559.1) and gorilla (NC001643.1) were aligned using MEGA v4 (Tamura et al. 2007). Two fragments of 96 and 20bp in the HV2 and MCR_F, respectively appear to have been historically deleted from the mitochondrial genome of orangutans but are present in both humans and chimpanzees. The proportion of variable sites (PVS), consisting of both indels and substitutions between species, was then calculated for the four MCR sub-domains and the two flanking regions using the program DnaSP v5 (Librado & Rozas 2009). The average number of numts per nucleotide position (numts/site) was estimated for each region. Regression analysis was used to compare the relationship between PVS and numts/site in order to test the effect of sequence variation on the number of detected numts.

Additionally, an index of DNA complexity was calculated by dividing the size in base pairs of each region by the number of base pairs considered to be part of nucleotide repeat blocks. Such blocks were determined by the program MSATFINDER v2.0 (Thurston & Field 2005) and defined as stretches of at least 5 tandem repeats of mononucleotides or at least 3 tandem repeats of longer motifs (2 to 6 nucleotides). Numt abundance was calculated as the number of numts partially or entirely derived from a particular region weighted by the size of the region. The relationship between DNA complexity and numt abundance was assessed through regression analysis in order to assess the effect of potential mutational hotspots in repetitive blocks (low complexity) on the ability to detect numts from different sub-domains.

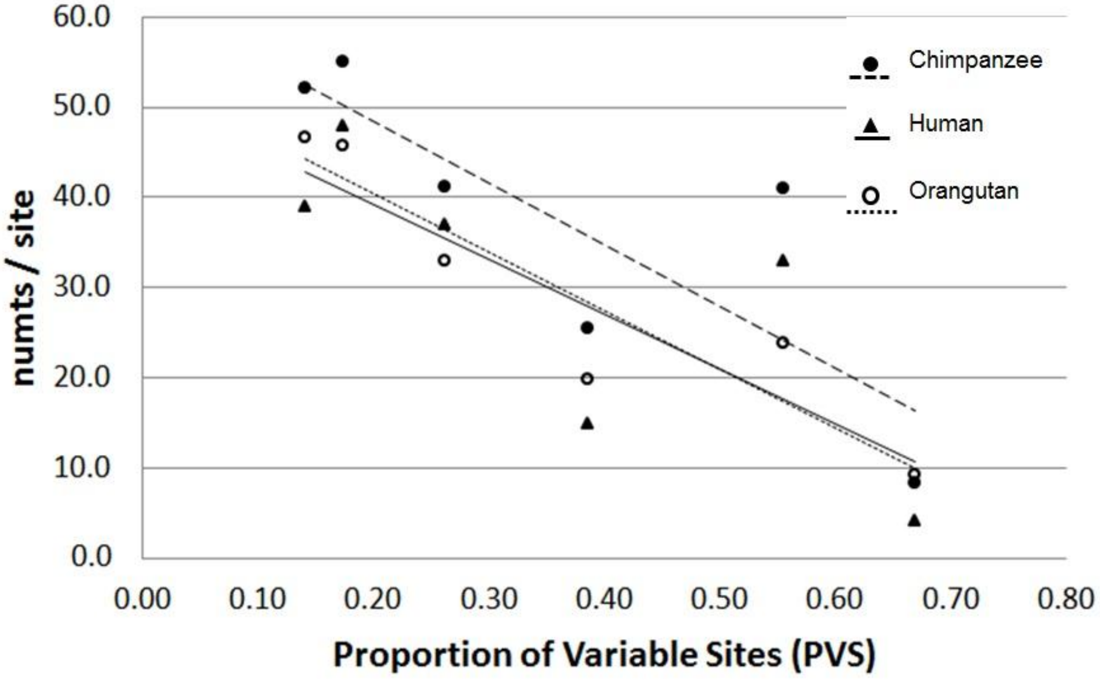
Insertion rate of MCR numts in the Hominoidea. The presence of human, chimpanzee and orangutan MCR numts in other Hominoidea and an outgroup macaque (rheMac2, Jan 2006) was first determined by genomic BLAT surveys (Kent 2002) of reference genomic databases or by BLAST analyses of trace files and shot-gun genomic reads from the white-cheeked crested gibbon (*Nomascus leucogenys*, ADFV00000000; September 2010) and western lowland gorilla (*Gorilla gorilla gorilla*, CABD00000000, November 2009). In cases where genomic sequences from gorilla and gibbon were not available or the location of orthologous regions was ambiguous, the presence/absence of a given MCR numt in these species was determined by cross-species PCR amplification of genomic DNA from western lowland gorilla or white-handed gibbon (*Hylobates lar*) using primers specific to the numt flanks (Appendix A). The probable age of each numt was then deduced by mapping the first appearance of a given insertion to the relevant inter-nodal position in the reference Hominoidea phylogeny, as proposed by Goodman et al. (1998). According to this phylogeny, the Cercopithecoidea (Old World monkeys including macaque) diverged from Hominoidea around 25Ma ago. The lineage leading to the gibbon then diverged 18Ma ago, followed by the divergence of *Pongo* (orangutan) 14Ma ago, *Gorilla* 7Ma ago and then the separation of the two terminal taxa *Homo* (human) and *Pan* (chimpanzee and bonobo) around 6Ma ago.

Figure 1.1. Absolute number of numts per site in the MCR and 500bp flanking regions of human, chimpanzee and orangutan. Sub-domains within the MCR are the first Hyper-variable region (HV1), the conserved central domain (CCD), the second Hyper-variable region (HV2) and the terminal sub-domain adjacent to the phenylalanine tRNA (MCR$_F$). The two 500bp flanking regions, MT$_P$ and MT$_F$, begin with the proline tRNA and phenylalanine tRNA, respectively.



The insertion rate of MCR numt loci was estimated as the total number of fragments that first appeared in a given inter-nodal region divided by the age difference between nodes. We did not attempt to conduct a rigorous distinction between independent mitochondrial translocations and post-integration duplications owing to the difficulty of unambiguously differentiating the two. However, several duplication events could be confirmed in cases where multiple numts exhibited the same boundaries and identity along their flanking regions (e.g. see panY8000 series in Appendix A).

Figure 1.2. Relationship between the proportion of variable sites (PVS) in the four MCR sub-domains and the two flanking regions and the average number of numts per nucleotide position (numts/site) in the human, chimpanzee and orangutan genomes. Regression equations are y = -60.933x + 51.497, y = -68.296x + 62.066 and y = -65.06x + 53.418 in human, chimpanzee and orangutan, respectively.

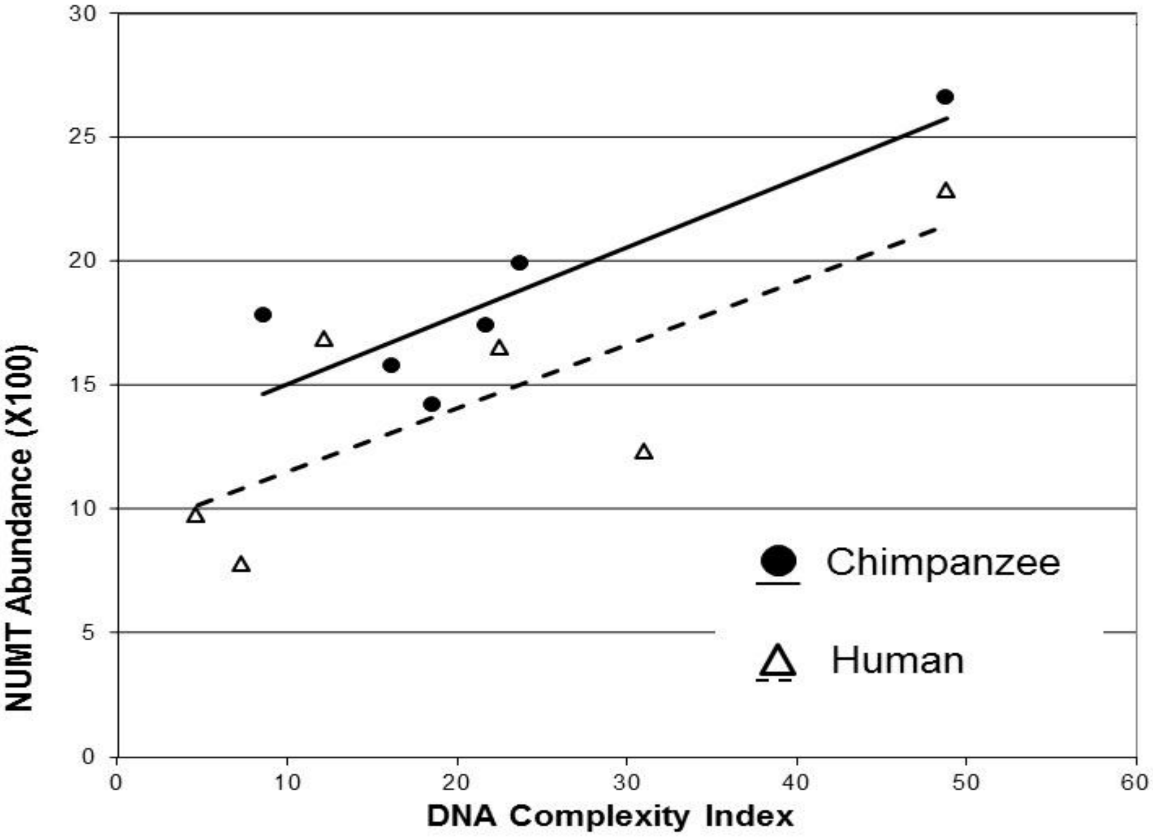


Results

BLAST searches recovered a total of 97 human, 122 chimpanzee and 100 orangutan putative MCR numt loci. There was a pronounced deficit in numts originating from the HV2 and MCR_F relative to the other two sub-domains (Figure 1.1). The relative proportion of numts by MCR sub-domain was similar in all three great ape taxa.

Structure of mtDNA sequence and the number of traceable numts. There was a negative relationship between the proportion of variable sites (PVS) and the number of numts/site in all three genomes (Figure 1.2). This relationship was highly significant for chimpanzees (Pearson=-0.83; d.f.=4; p=0.041; R^2=0.69) and orangutans (Pearson=-0.93; d.f.=4; p=0.007; R^2 = 0.87), but only marginally significant for humans (Pearson=-0.78; d.f.=4; p=0.063; R^2 = 0.62). On the other hand, there was a positive relationship between DNA complexity and numt abundance in humans, chimpanzees and orangutans (Figure 1.3). However, this relationship was only significant for chimpanzees (Pearson=0.87; d.f.=4; p=0.025; R^2=0.75) and marginally significant for humans (Pearson=0.78; d.f.=4; p=0.065; R^2=0.62).

Figure 1.3. Relationship between DNA complexity and numt abundance in humans and chimpanzees. Regression equations in chimpanzees and humans are y = 0.0028x + 0.123 and y = 0.0026x + 0.089, respectively.



Insertion rate of MCR numts. Genomic database surveys and cross-species PCR assays succeeded in placing the origin of 62 MCR numts in the hominoid phylogeny along with 22 additional numts derived from the two flanking regions MT_P (12) and MT_F (10) (Figure 1.4). MCR translocations include eight that originated prior to the divergence of orangutans, 25 specific to chimpanzees, 20 specific to orangutan and two specific to humans (see Appendices A and B for detailed information). Presence/absence status of five additional candidate numts could not be unambiguously determined in macaque due to gaps in the reference genome database or chromosomal deletions containing the target region. From these data, we estimated an average rate of insertion of 1.38 MCR numts per Ma in the hominoid genome, although this is likely to be slightly biased as numts in the lineages of gibbon and gorilla were missed. Different rates were found among taxa, with an outstandingly high rate in chimpanzee (4.17 numts/Ma) that contrasts with those in human (0.33 numt/Ma; the sister taxon) and orangutan (1.43/Ma).

Figure 1.4 a. Phylogeny of Hominoidea and macaque showing the number of MCR numts inserted during particular internodal time periods, the insertion rate (numts/Ma) and the sequence identity (%) with the mitochondrial query sequence. The numt family panY8000 was excluded from calculations of identity in chimpanzee since they are known to be duplications of an ancient numt and therefore do not represent the sequence identity in the chimpanzee lineage.

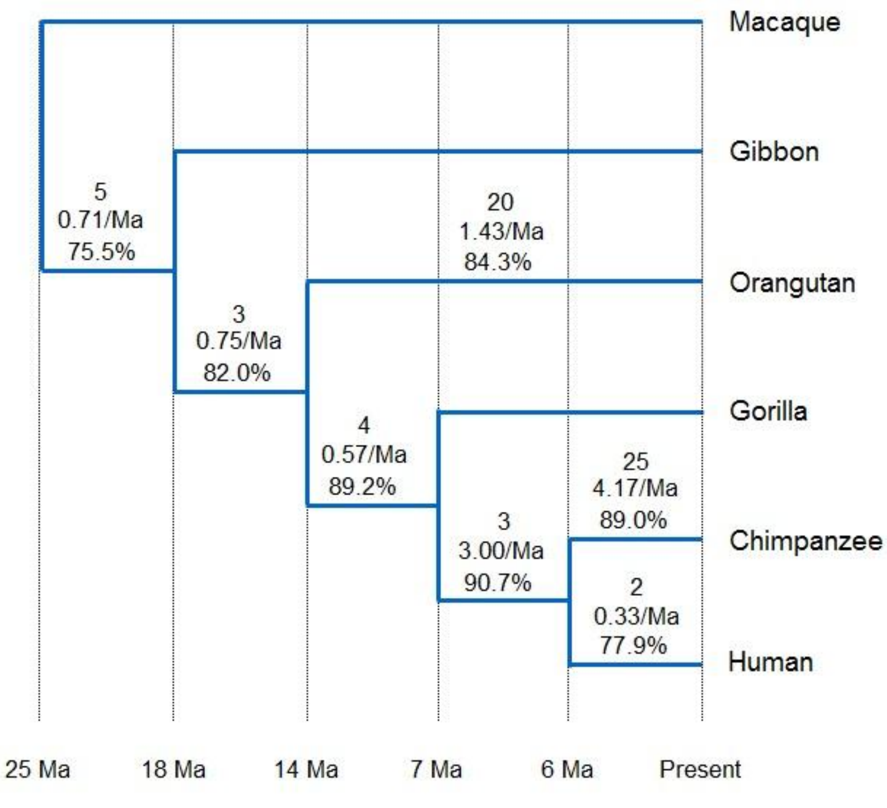
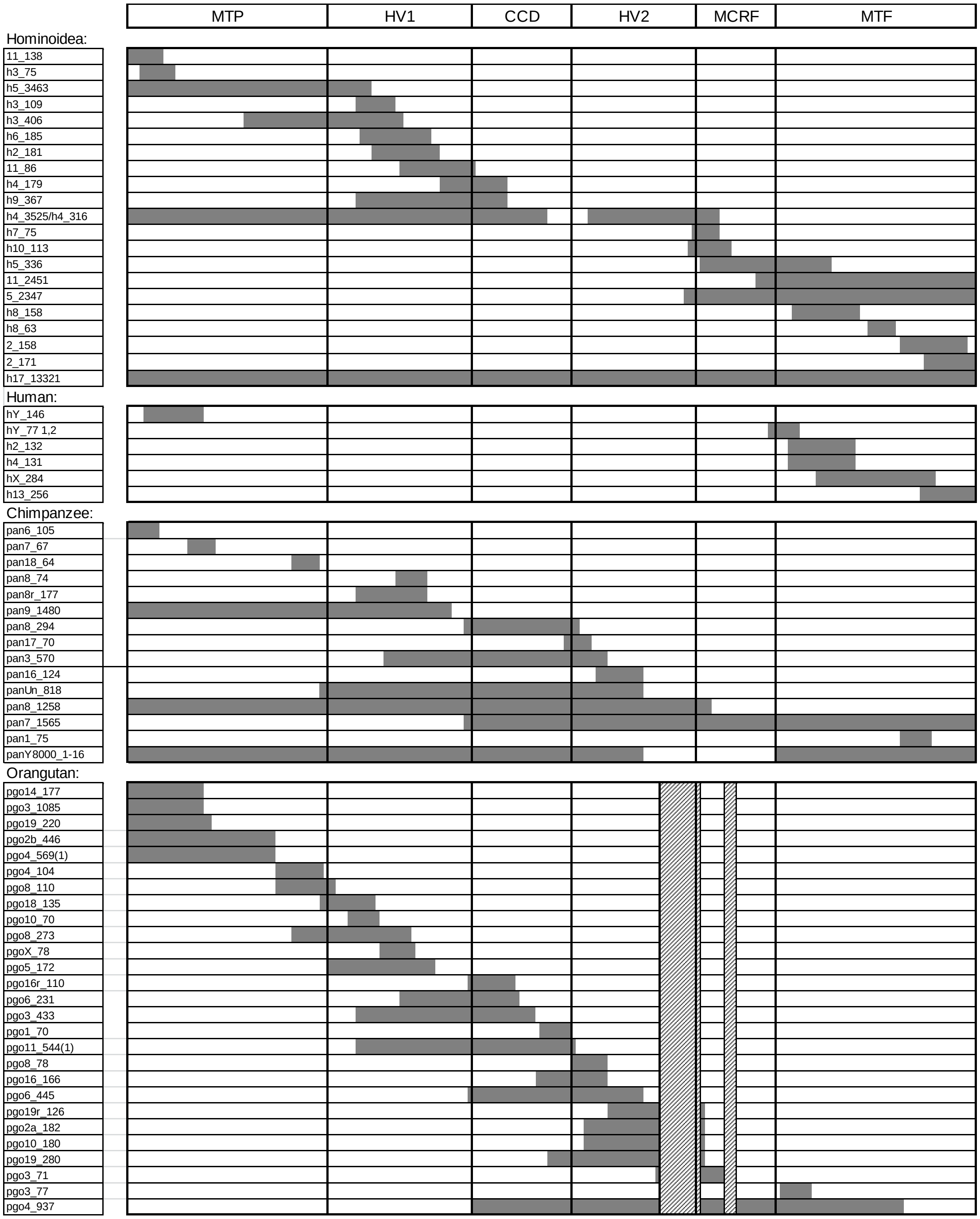


Figure 1.4 b. Hominoidea-specific numts derived from the region containing the MCR (HV1, CCD, HV2 and MCR_F) and 500bp of the flanking regions (MT_P and MT_F). They are organized in four groups depending on whether they are shared by multiple taxa (Hominoidea) or taxon-specific (human, chimpanzee or orangutan). Relative size and region of mitochondrial origin are depicted by gray boxes. Dashed boxes represent regions absent from the orangutan mitochondrial genome. See Appendix B for insertion time and specific chromosomal location of each numt.



Under-representation of the MCR sub-domains with largest sequence variation (HV2 and MCR_F) relative to other sub-domains was not observed in hominoid-specific numts, meaning that the overall numt deficit in those sub-domains is mostly determined by older numts. In general, sequence identity between mitochondrial sequences and their numt copies steadily decreased with numt age from nearly 90% in the human-chimpanzee ancestor to 75% in the hominoid ancestor, although this trend did not hold true for humans, where the two species-specific numts exhibited an identity of only 78% to one another (Figure 1.4 a).

Although we did not intend to make a rigorous distinction between direct integrations of mitochondrial fragments and duplications of previous integrations, we found multiple cases of recent MCR numt duplications nested in larger duplications of chromosomal fragments, interestingly all located in the Y chromosome. These comprise the two human-specific MCR numts (hY_77 1 and 2), which exhibit identical size, sequence and high identity along both flanks. Likewise, 15 of the 26 chimpanzee-specific numts were nested in chromosomal duplications; all located the Y chromosome (panY8000). They share identities of over 88% with one another and are derived from an ancient mitochondrial integration of ~8000bp that inserted over 25Ma ago in the Hominoidea ancestor. Altogether, panY8000 numts accounts for over 1.2 x 10^5 bp of mitochondrial sequences in the chimpanzee nuclear genome.

Discussion

The resulting list of MCR numts from our BLAST search recovered 40 human and 34 chimpanzee numts previously reported for these taxa (Mourier et al. 2001, Hazkani-Covo & Graur 2007, Lascaro et al. 2008, Ricchetti et al 2004, Zischler 1998, MITOMAP 2008), along with a large number of loci reported here for the first time including 7 in human and 23 in chimpanzee. Our study also identified 27 numt loci exclusive of the orangutan genome and provides the first comprehensive report of MCR numts in this taxon. The availability of three previous studies of human numts (Mourier et al. 2001; Hazkani-Covo & Graur 2007; Lascaro et al. 2008), enabled us to make a comparison with our own search and address the relevance of numt size and identity on the ability to detect extant numts. In particular for hominoid-specific MCR numts present in humans, numts found here and in one or more previous studies share on average 87 to 88% of sequence identity with their current mitochondrial genomes (Table 1). However, larger numts were more easily detected and are therefore more prevalent in earlier reports than smaller numts. In contrast, numts reported herein for the first time exhibit substantially smaller values of both average size (121bp) and sequence identity (73%) than previously reported numts. Overall, our results show that our search strategy recovered not only all previously described numts but also proved effective in uncovering additional numts with relatively small size and sequence identity. Although relaxing the search parameters in a BLAST survey is expected to increase the number of spurious associations, it is certainly useful in detecting real numts whose authenticity can be proved by amplifying these loci using primers targeted against the nuclear flanks in order to establish presence/absence comparisons between taxa. Such approach to numt detection could also prove useful in identification of cryptic numts in other species as one step in avoiding their inadvertent incorporation in future studies.

Table 1. Number of Hominoidea-specific numts derived from MCR and 500bp flanking regions reported in previous searches in the human genome. Table shows the number of loci (n) as well as size and identity for numts reported by previous studies: Mourier et al. 2001 (a); Hazkani-Covo & Graur 2007 (b); Lascaro et al. 2008 (c).

Previous studies	n	Average Numt size (bp)	% Identity
a, b, c	9	2954.8	88.4%
a, b or b, c	8	200.9	88.0%
b	4	140.8	87.0%
Newly reported	7	120.6	73.0%

A relative deficit in the number of numts derived from HV2 and MCR_F of the MCR was observed in humans, chimpanzees and orangutans. The strong negative relationship between mitochondrial PVS and the number of numts/site supports the hypothesis that elevated rate of sequence evolution in the mtDNA erodes sequence identity and leads to an apparent deficit in the amount of mitochondrial sequences detected in the nuclear genome. The positive relationship between complexity and numt abundance in humans and chimpanzees indicates that the loss of sequence identity and our ability to detect numts can be partially explained by elevated mutation rates in low complexity regions of the mitochondrial genome (Bodenteich et al. 1992; Sbisà et al. 1997; Zardoya & Meyer 1998). In other words, numts are less likely to be detected if they contain regions of the mitochondrial genome of higher mutation rate and repetitive sequence content. This then might also explain the apparent deficit of MCR numts relative to insertions from other parts of the mitochondrial genome (Mourier et al. 2001).

Additionally, recent deletions in mitochondrial genomes used as query sequences, such as the case of two fragments deleted from the mtDNA of orangutans, may also result in limited detection of numts that originated prior to the mitochondrial deletion. Overall, future search strategies should experiment with using query sequences either derived from consensus sequences of multiple taxa or from an inferred ancestral mitochondrial sequence of Hominoidea in order to determine whether more divergent nuclear translocations or translocations derived from regions no longer present in the mitochondrial genome can be detected in this way.

Several pieces of evidence point to the possibility that previous analyses based on humans have underestimated the rate of insertion in other great apes. Firstly, humans are known to have reduced genetic diversity relative to other apes due to a past population bottleneck (Zhao et al. 2000; Kaessmann et al. 2001; Mathews et al. 2003) which, as shown here, might have led to a numt deficit relative to other apes. Secondly, BLAST surveys of genomic databases based on a single individual are likely to underestimate the frequency of recent insertions that have not yet become fixed in the species (Schmitz et al. 2005). Lastly, although our search identified previously unreported numts in the hominoid genome, either partially or entirely derived from the MCR, our estimated rate is still likely to be a conservative value due to our deliberate exclusion of numt hits shorter than 50bp. Taken together, findings from this study provide strong evidence that MCR numts may be generally underestimated in most genomic surveys of existing genomic databases of great apes. In order to combat this problem, we recommend incorporating as many individual genomes as become available in future genomic surveys, combined with previous suggestions such as relaxing parameters in BLAST searches and the use of alternative query sequences.

There was also substantial variation in the rate of insertion among different taxa included in this study. Such differences are unlikely to result from a systematic bias in the BLAST methods used here since these were the same in all three taxa. Rate heterogeneity among lineages cannot be attributed to a bias introduced by gaps in genome projects since the slowest rate of insertion was found in the human genome whose sequencing project is the most comprehensive. The outstanding difference between humans and chimpanzees, despite the relatively recent divergence of these two taxa is also in agreement with previous reports of a larger number of numts in the chimpanzee genome (Hazkani-Covo & Graur 2007; Hazkani-Covo 2009). The observed deficit in humans is consistent with the historically low levels of genetic variation in humans, presumed to have arisen as a result of a historical bottleneck in this species as evidenced by high levels of neutral genetic variation still present in other apes (Kaessmann et al. 2001; Gherman et al. 2007).

Our results also indicate that the deficit of numts from particular MCR sub-domains is mostly due to loss of mitochondrial identity in numts inserted before the diversification of Hominoidea. This is supported by the fact that no obvious under-representation of HV2 and MCR_F was detected in numts originating during the last 25Ma, despite the relatively rapid divergence of these sub-domains. Also, the fact that the relative abundance of numts between MCR sub-domains exhibits similar patterns in all three hominoids studied here suggests a long history of mitochondrial migration into the nucleus prior to the divergence of hominoids.

In contrast, recent numts exhibit an increased sequence identity with current mitochondrial genomes. An exception to this was the relatively low identity between the human-specific numts and the mitochondrial genome but this is likely due to a sampling error since only two MCR numts were detected in humans and they are identical to one another. Overall, the high resemblance between the sequence of mitochondrial and nuclear copies may be potentially problematic and lead to misidentification of recent numts as mitochondrial sequences in population genetic studies (Jensen-Seaman et al. 2004). In these cases, inventories of species-specific numts characterized through either BLAST surveys of existing genomic databases or cross-species PCR assays will help identify instances of numt contamination and ensure that mitochondrial sequence databases are error-free.

Our findings showed a recent accumulation of numt duplications in the Y chromosome of the Hominoidea. The majority of species-specific MCR numts found in chimpanzees and the only two found in humans were located in the Y chromosome, nested within recent duplications of larger chromosomal segments. Unfortunately, comparisons with orangutan are not possible since sequence data of the Y chromosome are not available in the current version of the genome project. A concentration of numts in the Y chromosome despite its small size is also supported by data from an early draft of the human genome where an excess of human-specific numts relative to the chromosomal size was found in this chromosome (Ricchetti et al. 2004). The Y chromosome is known to have unusually repetitive content, whose reduced gene density and relaxed functional constraints provide the basis for numerous chromosomal changes including deletions, integrations and duplications (Foote et al. 1992; Tilford et al. 2001). Moreover, the greater number of cell divisions in the male germ line can also facilitate a vast accumulation of chromosomal rearrangements (Erlandsson et al. 2000). Future sequencing of multiple conspecific genomes and completion of other ongoing genome projects may shed light on whether the observed concentration of recent numts in the Y chromosome is common to other primates or varies between populations and subspecies. If that is the case, then duplications of chromosomal fragments may prove useful as cytogenetic markers in future population genetic studies.

CHAPTER 2

NUCLEOTIDE COMPOSITION, SEQUENCE EVOLUTION AND MECHANISMS OF INSERTION OF NUCLEAR COPIES OF MITOCHONDRIAL DNA IN GREAT APES (HOMINOIDEA)

Abstract

The widespread distribution of copies of mitochondrial DNA (mtDNA) in the nuclear genome of great apes, also called numts, provides an unparalleled opportunity to compare the evolution of mitochondrial sequences and their paralogous copies in the nuclear genome. While it is generally acknowledged that patterns of nucleotide substitution, sequence composition and selection will differ between mitochondrial and nuclear environments, comparative data are lacking. Similarly, knowledge is also lacking on the potential mechanisms of nuclear integration and their relative importance across multiple taxa. Here, we built on a large dataset (n=83) of great ape-specific numts and their mitochondrial paralogs to: (1) quantify differences in transition/transversion ratios and GC content between mitochondrial sequences and their corresponding numts; (2) explore the extent to which sequence stability of human mitochondrial genes might affect their identity to nuclear copies and the ability to detect the latter in genomic databases (3) examine the relative importance of different mechanisms of numt insertion in great ape genomes. In general, transition/transversion ratios differed significantly between both mitochondrial and nuclear sequences and between numts derived from coding and non-coding mitochondrial regions. The previously documented upward bias in the GC content of the primate mitochondrial genome was confirmed and the extent of this bias relative to the corresponding numt sequences increased with numt age. Conserved human mitochondrial genes maintain a higher identity with nuclear copies and because of this, appear to be over-represented in human numt databases. Lastly, comparison of alternative mechanisms of numt insertion revealed that non-homologous end joining repair is the most likely mechanism of numt integration in great apes.

Introduction

Comparative analysis of whole genomes has unveiled a great abundance of mitochondrial DNA (mtDNA) sequences inserted in the nuclear genome (numts) of many eukaryotes (Bensasson et al. 2001; Hazkani-Covo 2010). In great apes, numts are particularly widespread at levels that apparently surpass those found in other mammals (Clifford et al. 2004; Jensen-Seaman et al. 2004; Richly & Leister 2004; Triant & DeWoody 2007; Anthony et al. 2007a). This poses a real problem for evolutionary studies of these species due to the high risk of incorporating numts into downstream mitochondrial analyses (Thalmann et al. 2004; Jensen-Seaman et al. 2004; Anthony et al. 2007a). However, the prevalence of numts in great apes (Hominoidea) also provides an unrivaled opportunity to study how nuclear integration affects the molecular evolutionary dynamics of mitochondrial sequences. This can best be accomplished through a systematic comparison of the sequence properties of a large suite of mitochondrial sequences with their corresponding nuclear paralogs taken at different time points during the evolution of great apes.

Once mitochondrial fragments escape to the nucleus they become non-functional sequences and as such are released from selective mitochondrial constraints. As a result, it is thought that these integrations mutate at rates that resemble non-coding nuclear loci, which are around one order of magnitude slower than the mitochondrial average (Brown et al. 1982; Graur & Li 2000; Haag-Liautard et al. 2008). For this reason, numts can be considered to be "molecular fossils" of ancient mitochondrial lineages that retain the sequence composition of mitochondrial genomes at the time of insertion (Perna & Kocher 1996; Bensasson et al. 2001; Zischler et al. 1995b).

Of particular interest to the present study is the observation that the mitochondrial genome of Old World primates has an intrinsic upward elevational bias in GC content (Schmitz et al. 2002), hypothesized to have arisen as a result of lineage-specific mutational pressure (Schmitz et al. 2002; Gibson et al. 2004). Following translocation, great ape numts escape from this GC mutational pressure and exhibit a GC content that is lower than their current mitochondrial counterparts (Schmitz et al. 2002). This difference in GC content between nuclear and mitochondrial copies is likely to be greater in numts that transferred earlier into a nuclear environment than those that have only recently been translocated. We therefore predict that a positive relationship exists between the age of the insertion and the magnitude of the difference in the GC content and that such relationship may be used to estimate the rate at which GC content increases in the mitochondrial genome. However, such relationship remains to be explored.

Previous studies of sequence evolution in primate pseudogenes have also shown that high regional GC content and the presence of GC dinucleotides may affect the mutational dynamics of neighboring nucleotide positions by boosting the number of transitions, resulting in an elevated proportion of transitions (T_s) over transversions (T_v), or T_s/T_v ratio (Bulmer et al. 1986, Hess et al. 1994). Since primate numts have a tendency to be GC rich, it is thus possible that they also have intrinsically large T_s/T_v ratios. The mitochondrial genome also exhibits a strong transition-biased nucleotide substitution pattern and elevated substitution rate that frequently results in saturation in the number of transitions and underestimation of T_s/T_v ratios (Arctander 1995; Lopez et al. 1997; Purvis et al. 1997; Yang & Yoder 1999). In contrast, T_s/T_v ratios in numts are thought to be less sensitive to saturation given the low substitution rate of the nuclear genome. Only a few studies, mostly based on a limited number of numts derived from confined mitochondrial regions have addressed differences in T_s/T_v ratios between mitochondrial and nuclear copies (Lopez et al. 1997; Zischler et al. 1998; Mundy et al. 2000; Schmitz et al. 2005).

With respect to rates of translocation of different mitochondrial fragments into the nuclear genome, early numt searches in humans and other great apes have shown that numts originating from the non-coding mitochondrial control region (MCR) are poorly represented in the nuclear genome (Mourier et al. 2001). Two alternative hypotheses have been proposed: First, a transfer of genetic material to the nucleus might be preferentially mediated by mRNA thus explaining the deficit of numts from non-coding regions such as MCR. Although such a mechanism has been demonstrated in plants (Henze & Martin) it remains to be shown in animals. Secondly, the apparent deficit in the number of numts from the MCR might be caused by the elevated mutation rate of this region and its subsequent rapid loss of identity with the translocated nuclear copies (Sbisà et al. 1997; Pesole et al. 1999; Mourier et al. 2001; Soto-Calderón et al. *in review*). In contrast to the MCR, we predict that more conserved mitochondrial genes should maintain a relatively high identity with nuclear copies. This effect may then lead to an apparent elevation in the number of detectable nuclear copies of such genes relative to nuclear copies of more variable mitochondrial genes.

In addition to studies of the nucleotide properties of numts, there has been considerable interest in the mechanisms by which mtDNA colonizes the nuclear genome. Three potential mechanisms of insertion have been proposed. The first is based on Non-Homologous End-

Joining Repair (NHEJR), where base complementarity of one to seven bases (or microhomologies) between two sequences can often facilitate recognition between the mitochondrial fragment and its nuclear background (Blanchard & Schmidt 1996; Jackson 2002). Although microhomologies are not a necessary requirement for NHEJR to occur, their presence is likely to be a signature of this mechanism. Secondly, *trans*-replication slippage has been postulated to play a role in integration of mitochondrial fragments into the nuclear genome (see Chen et al. 2005). This mechanism, first described by Chen et al. (2005), relies on a complex sequence of events whose outcome is distinguished by the presence of a nucleotide motif in both endings of the mitochondrial insertion and the integration site. Lastly, it has also been suggested that Transposable Elements (TEs) might mediate integration of mtDNA into the nucleus since these two elements are frequently found in close proximity (Farrelly & Butow 1983; Ricchetti et al. 1999; 2004; Lascaro et al. 2008). For instance, Mishmar et al. (2004) found that members of a particular family of TEs called Long INterspersed Elements (LINEs) are preferentially integrated within 150bp of numts in the human genome, suggesting a non-random association of TEs and numts. Conversely, more recent studies have revealed an apparent deficit of TEs within 200bp of numt loci (Gherman et al. 2007; Jensen-Seaman et al. 2009). Interestingly, a particular group of LINEs named LINE-type 1 (LINE-1 or L1) could potentially facilitate the retrotransposition of flanking non-LINE elements resulting in simultaneous duplication of both a LINE-1 and the flank (Moran et al. 1999; Pickeral et al. 2000; Goodier et al. 2000; Deininger et al. 2003). However, the contribution of such a mechanism to duplication of flanking numts and its frequency relative to other hypothesized modes of numt dissemination has not yet been systematically compared.

Here, we make use of 83 dated numts across the entire Hominoidea phylogeny to assess differences in nucleotide composition and patterns of substitution between mitochondrial and numt sequences of various ages and mitochondrial regions. In order to test the hypothesis that sequence stability in mitochondrial genes is positively related to numt prevalence, we also compare the proportion of variable sites in 15 mitochondrial genes in humans (Ingman & Gyllensten 2006) to the prevalence of their nuclear pseudogenes (Triant & deWoody 2007). Lastly, we examine the flanking sequences of all 83 great ape numts to evaluate the relative importance of NHEJR, *trans*-replication slippage and TEs in the numt insertion process.

Materials and Methods

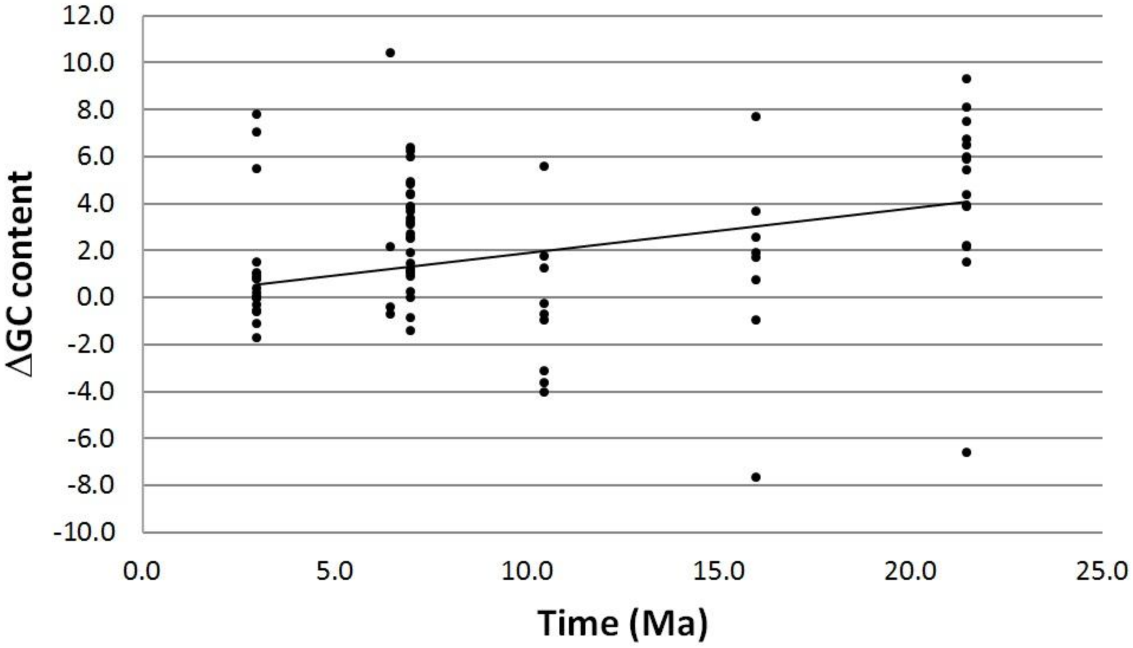
Selection of numt loci and estimation of insertion time. We assembled a database of 83 hominoid numt loci exclusively present in great apes comprising 47 derived from MCR and 36 from other mitochondrial regions (non-MCR). Exclusive presence of a numt in Hominoidea was inferred by verifying its absence from the macaque genome (rheMac2, Jan 2006), a member of the sister group Cercopithecoidea, using the BLAT tool in the UCSC Genome Browser (Kent et al. 2002). The numts used in this study were either retrieved from previous genomic BLAST searches in the human, chimpanzee, and orangutan reference genomes (Soto-Calderón et al. *in review*), or reported in other studies (Hazkani-Covo et al. 2007; Jensen-Seaman et al. 2009) (Appendix C). In order to estimate the approximate age of each numt, presence/absence was determined in the reference genomes of human, chimpanzee and orangutan using the BLAT tool. Presence/absence in gibbon and gorilla was determined through BLAST searches against the partial genomic databases of *Nomascus leucogenys* (ADFV00000000; September 2010) and *Gorilla gorilla gorilla* (CABD00000000, November 2009), respectively. Whenever orthologous regions were absent from a reference genomic database, locus-specific primer pairs were designed from human and chimpanzee alignments of the corresponding region and used to amplify the target locus from gorilla and/or gibbon genomic DNA. These amplified PCR products were sequenced using the Big-Dye version 1.1 (ABI) and run on an ABI 3100 genetic analyzer (Appendix D). Based on its presence/absence pattern in the reference genomes, the approximate time of insertion of each numt was then inferred by mapping the time of its first appearance onto an internodal time period in a reference Hominoidea phylogeny (see Figure 2.1; Goodman et al. 1998). In this way, divergence times were defined as follows: Human/Chimpanzee - 6.5Ma; Human/Gorilla and Chimpanzee/Gorilla - 10.5Ma; Human/Orangutan, Chimpanzee/Orangutan and Gorilla/Orangutan - 16Ma and; Human/Gibbon, Chimpanzee/Gibbon, Gorilla/Gibbon and Orangutan/Gibbon - 21.5Ma.

Patterns of nucleotide substitution in Hominoidea numts and their mitochondrial counterparts. Sequences were aligned using the ClustalW algorithm (Larkin et al. 2007) implemented in the program MEGA v4.0 (Tamura et al. 2007). The program DnaSP 5.10 (Librado & Rozas 2009) was used to estimate the average GC content in each numt and its mitochondrial paralog. Due to the GC bias in contemporary mitochondrial genomes, a one-tailed paired t-test was then used to test for an excess in the average GC content (ΔGC) in mitochondrial relative to nuclear sequences. Regression analysis and Spearman's rho correlation tests were used to assess the relationship between pair-wise ΔGC differences and numt-mitochondrial divergence time. The program PAUP v4.0b10 (Swofford 2002) was used to estimate the average observed T_s/T_v ratio for each numt present in at least two taxa (n=36) and for their corresponding mitochondrial sequences. The relationship between the natural logarithm of T_s/T_v ratios and GC content was assessed using a Pearson's rho correlation tests. The T_s/T_v ratios obtained from pair-wise sequence comparisons between taxa were estimated for MCR and non-MCR numts and differences were assessed using a two-tailed paired t-test.

Selection on mitochondrial genes and the apparent prevalence of nuclear copies. We tested the correlation between the proportion of variable sites (PVS) of 15 human mitochondrial genes with the number of numts derived from the same genes in the human genome. To do this, we made use of 100 mitochondrial genomes from human populations around the world available through the Human Mitochondrial Genome Database (Ingman & Gyllensten 2006; see Appendix E) and an inventory of human numts found through BLAST searches for the 13 protein-coding and the 2 rRNA mitochondrial genes in humans, as reported by Triant & deWoody (2007). PVS and a proportion of the average number of numts per nucleotide position (numts/site) were calculated as in the previous chapter for the 15 mitochondrial genes. The relationship between PVS and numts/site was assessed using a Pearson correlation test.

Mechanisms of numt insertion. We inferred the boundaries of the putative pre-integration site of each numt locus from the next most basal taxon in the phylogeny lacking the numt in question (Hazkani-Covo & Covo 2008; Jensen-Seaman et al. 2009). We then compared the termini of each numt with that of the sequence lacking the numt to survey for the presence of microhomologies on one or both flanks (see Jensen-Seaman et al. 2009). Presence of microhomologies in either flank was considered as indirect evidence of NHEJR. Presence of the same microhomology motif on both numt endings was assumed to be a signature of *trans*-replication slippage. In addition to these two mechanisms of direct mitochondrial integration, the potential role of transduction via TE-LINE1 in numt duplication was also assessed. A signature of this mechanism would be the adjacent location of a target numt and a TE that arose within the same inter-nodal period in the reference phylogeny. This association was determined through a search of TEs located within 500bp of each target numt and assessing their presence/absence in the reference genomes of human, chimpanzee and orangutan. TEs were tracked using the program REPEATMASKER-Open 3.0 (Smit et al. 1996-2007), available through the UCSC Genome Browser (Kent et al. 2002). Finally, we assessed whether there is any evidence that human numts are preferentially associated with one or more families of TE (i.e. Alu, LINE-1, MIR, LTR and LINE-2). To do this, we identified TEs integrated within the same or previous time period of a human numt located within 500bp and used a χ^2 test to compare their proportions with the distribution of TE families in the whole human genome (Deininger & Batzer 2002).

Figure 2.1. Relationship between time of numt insertion and the difference in GC content between numts and mtDNA. The divergence time was defined as the mid-point of the internodal time period basal to the node connecting the sequences under comparison.



Results

Differences in rates and patterns of nucleotide substitution between numts and their mitochondrial counterparts. A comparison between mitochondrial regions revealed that the GC content in the MCR (48.1%) of Hominoidea was 3.5% higher than the mitochondrial average (44.6%). Also, GC content was significantly lower in numts (43.31% ± 1.14) than the mitochondrial counterparts (45.50% ± 1.10; paired $t_0 = 5.96$; $p<0.001$; d.f. = 82). The ΔGC between the nuclear and the mitochondrial genome increased with species divergence time (Spearman=0.288; p=0.008) (Figure 2.1) at an approximate rate of 2% every 10Ma. The observed pair-wise T_s/T_v ratio in mitochondrial alignments decreased with time of divergence between taxa whereas in numts this ratio tends to increase with time. The pair-wise T_s/T_v ratio of MCR numts was significantly greater (paired $t_0 = 5.68$; $p<0.001$; d.f. = 15) than the T_s/T_v ratio of non-MCR numts (Figure 2.2). GC content and the average T_s/T_v ratio in numts showed a positive relationship, although this was not significant (Figure 2.3).

Table 2. Microhomologies between clusters of adjacent numts.

Cluster 1 h4_236 h4_60 h4_316 h4_3525 h4_1345	No microhomologies. GATTAAAATT TTAAAATTATAC No microhomologies. No microhomologies.

Cluster 2
8_68(1)
8_68(2)
	No microhomologies.

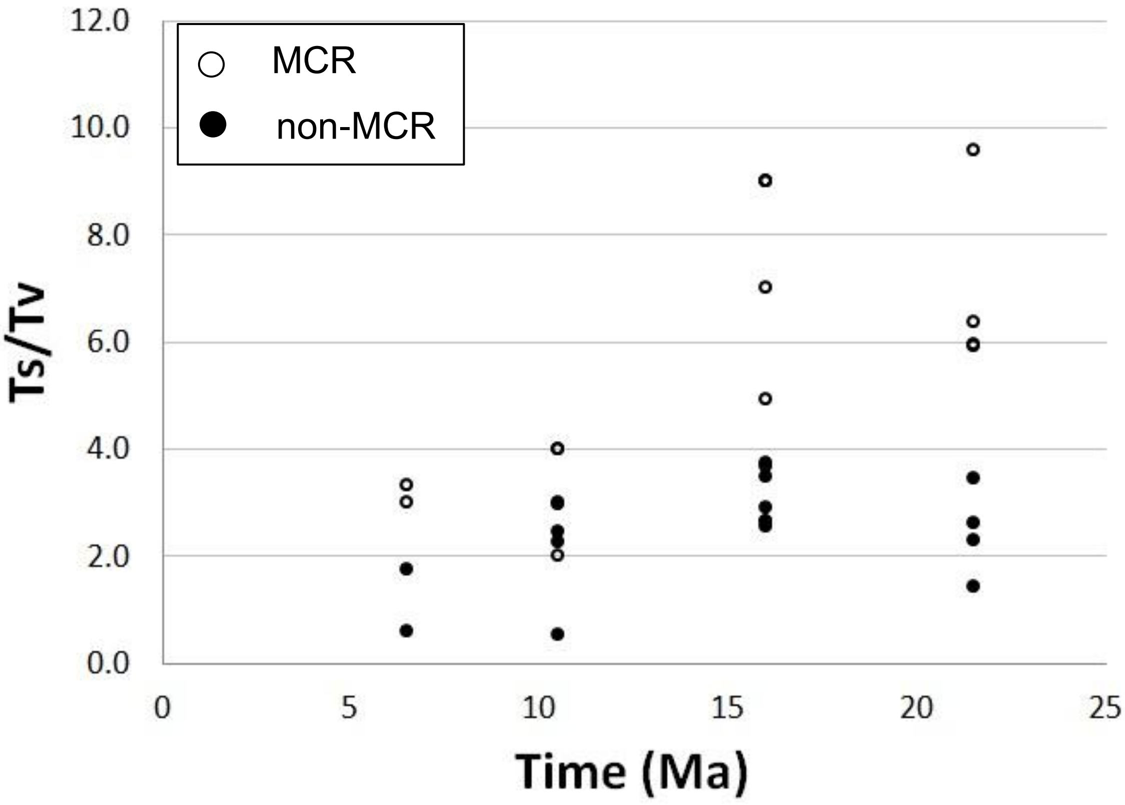
Cluster 3 h17_13321 17_232	CATATT TATTGA
Cluster 4 h3_109 3_76 3_136	TACCCC CCCTG……TCGGG GGGTG
Cluster 5 hX_749 hX_284 hX_554	AATAT TATTG………AATCATA TCATAACCC
Cluster 6 pan6_85 pan6_105	No microhomologies.

Cluster 7
pgo4_569(1)
pgo4_569(2)
	TTGATCCTGTTTCGTGTAGAAATAGGAGGTGTAGGGTTGTTAGAGCT
	GATCCTGTTTCGTGTAGAAATAGGAGGTGTAGGGTTGTTAGAGCTAG

Cluster 8
pgo11_544(1)
pgo11_544(2)
	GCCCACCCAGATAAAA
	CCACCCAGATAAAAAT

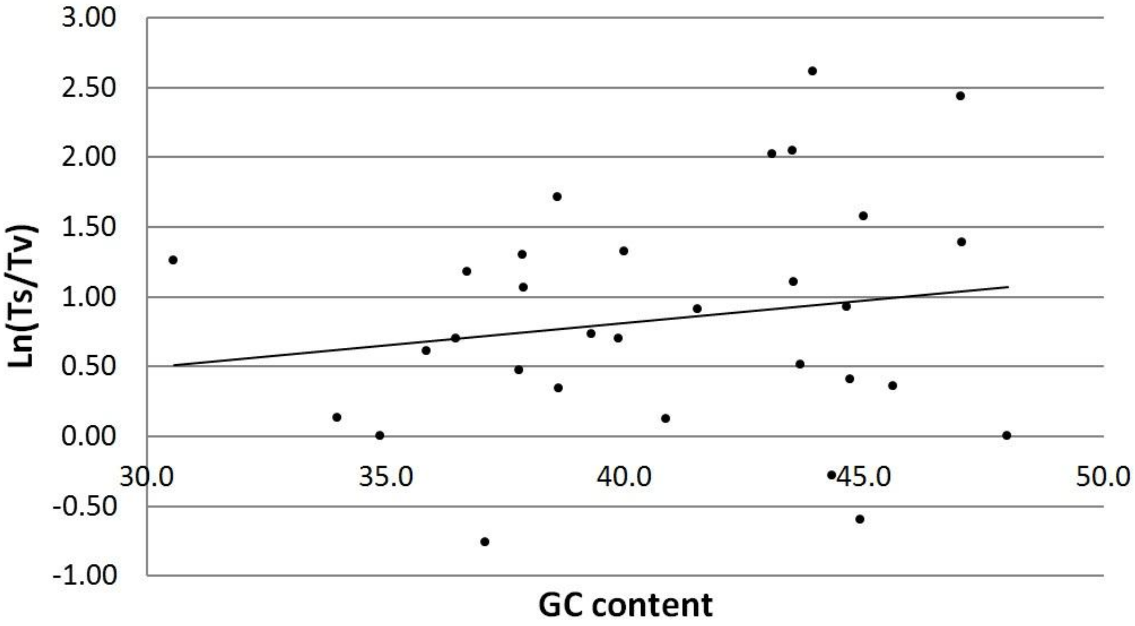
Selective patterns in mitochondrial protein coding genes and number of numts. Analysis of the 15 mitochondrial genes in humans revealed that the average number of numts derived from each nucleotide site decreased with the proportion of variable sites in each gene (Figure 2.4). This results in a negative logarithmic relationship between PVS and numts/site (Pearson = -0.64; p = 0.010; d.f. = 13). For instance, the gene 16S has the smallest PVS (0.030) and one of the greatest proportion of numts per site (23.6, whereas the gene ATP8 has a relatively high PVS value of 0.09 and a proportion of numts/site of only 6.64.

Figure 2.2. T_s/T_v ratios of MCR and non-MCR numts.



Mechanisms of numt insertion. Inspection of junction sites between numt termini and their corresponding flanking nuclear sequences revealed the presence of microhomologies in 75 (45%) out of 166 numt boundaries. The remaining numt junctions comprised 82 cases where microhomologies were not present and 9 cases where the numt boundary could not be identified due to gaps in reference sequences. Microhomology size ranged from 1 to 13 nucleotides, plus one special case where an unusual homology of 42 nucleotides was detected in an orangutan numt of over 1080 bp long (pgo3_1085). Perfect and imperfect microhomologies were present in 61 and 14 numt junctions, respectively. There was also an overall decline in the frequency of each microhomology class with size of motif. For numts where both junctions were characterized, 26% of all cases exhibited microhomologies in the two junctions, 47% in only one and 27% in none. In no instance was the same polynucleotide motif observed on both endings.

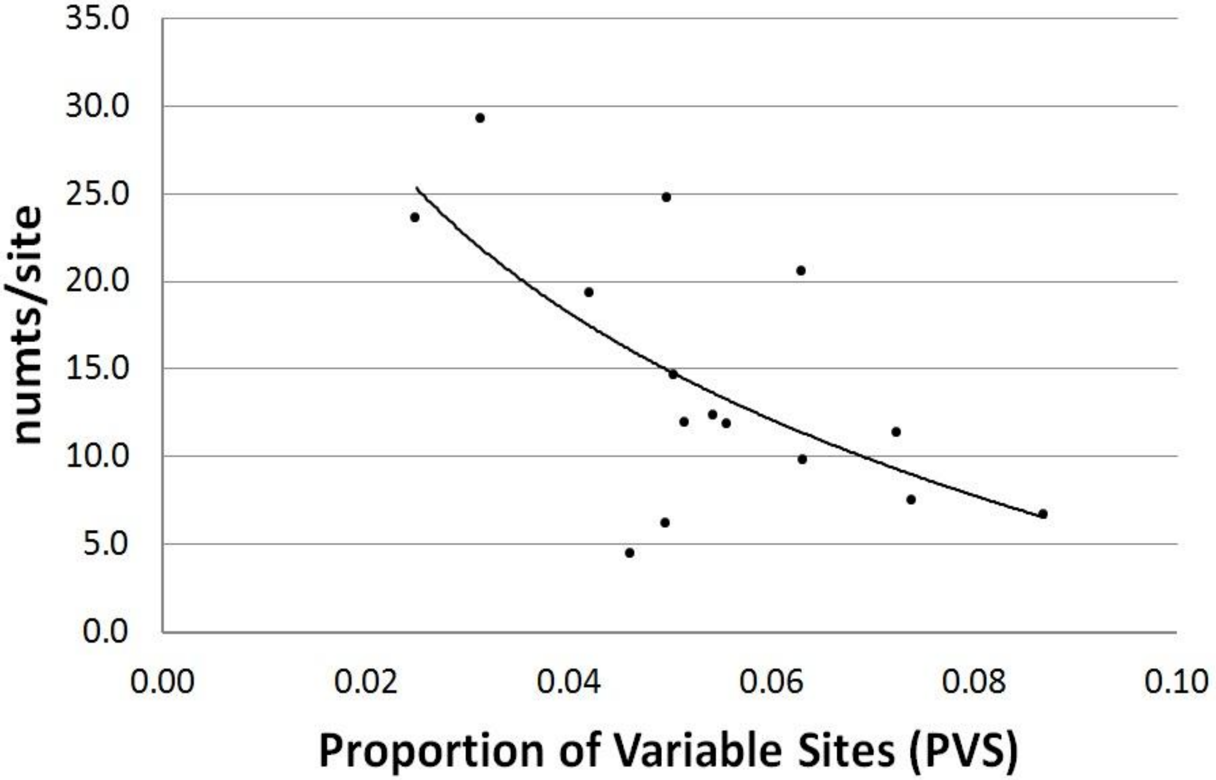
Figure 2.3. Relationship between the average pairwise T_s/T_v ratio in 33 numt loci present in at least three taxa and their GC content. y = 0.0321x - 0.4667; R² = 0.0287.



Microhomologies were also found between adjacent numt fragments inserted during the same inter-nodal period in the reference phylogeny (Appendix C; Table 2). A total of eight numt loci were made up of multiple tandemly-arranged numt fragments including six cases with microhomologies between contiguous mitochondrial insertions. Homology in these complex numts ranged from one to 45 nucleotides. Of these six cases, two numts were made of three fragments that displayed microhomologies in all boundaries between them. The most extensive microhomology was a stretch of 45 nucleotides present in both pgo4_569(1) and pgo4_569(2).

Inspection of nuclear flanking sequences 500bp either side of a numt yielded three cases where the time of insertion coincided with the insertion of nearby TEs. The first case is numt 2_592, inserted 18 - 14 Ma ago and located within 10bp of an array of one TE-Alu element and two consecutive TE-MIRs (mammalian-wide interspersed repeats). The second case is the numt h4_179, located 150bp from a TE-Alu inserted within the same time period 18 - 14 Ma ago. The intervening sequence corresponds to the remnant of a TE-LTR (Long Terminal Repeat) already present at the time of the insertion of the two other elements. The third case corresponds to the numt pgo2a_182 which is associated with a TE-LTR exclusively present in orangutan and separated by 460bp from an older TE-LTR. In all other surveys, TEs were either absent from the 500bp regions flanking a target numt or integrated at different time periods. Additionally, the frequency distribution of TE families (Alu, LINE-1, MIR, LTR and L2) in the 500bp neighboring human numts did not differ from the observed distribution in the whole human genome ($\chi^2 = 8.72$; p = 0.069; d.f. = 4), indicating that numt duplication is not associated with any particular TE activity.

Figure 2.4. Relationship between the proportion of variable sites (PVS) in 15 mitochondrial genes in humans and the average proportion of numts per site (numts/site) for a given gene size. The regression equation is y = -15.09ln(x) - 30.325.



Discussion

We combined multiple loci derived from coding and non-coding mitochondrial domains and show that the observed GC content of sequences in the mitochondrial genome exceeds that of their nuclear copies, in agreement with the hypothesis of genome-wide mutational pressure in the mitochondrial genome of the Hominoidea (Schmitz et al. 2002; Gibson et al. 2004). The accumulation of differences in GC content with divergence time indicates that mutational bias in mtDNA has been an ongoing process at least since the origin of diversification of the Hominoidea. Schmitz et al. (2002) previously found a GC bias in mitochondrial sequence composition in both synonymous and non-synonymous sites, consistent with a genome-wide directional mutational mechanism. Indeed, this bias is not limited to primates since a comparison of mammalian mitochondrial genomes has shown that all codon positions and rRNAs within the same DNA strand are affected by the same compositional changes (Gibson et al. 2004). However, this compositional bias can vary between phylogenetically related taxa suggesting a “switch” that can change mutational direction in one or another. The factors that affect this bias are not well understood but elevated levels of C and low levels of T in the L-strand are associated with the time spent as single strand during mitochondrial replication (Gibson et al. 2004), which in turn is correlated with mutation rates across the mitochondrial genome (Broughton & Reneau 2006). Thus, an excess of C, and therefore GC content together with a deficit of AT bases might be determined to some extent by variation in the mechanism of mitochondrial replication.

In numts, T_s/T_v ratios vary with the depth of divergence of the sequences under comparison and the genomic context in which these comparisons are made. The observed negative relationship between T_s/T_v ratios and divergence time of mtDNA but not numts is likely a consequence of the elevated substitution rate of the former. This leads to saturation in the number of transitions at higher levels of genetic divergence and underestimation of T_s/T_v ratios (Yang & Yoder 1999). Compared to mitochondria, the T_s/T_v ratio in numts increased slightly with time of divergence and remained within the range previously reported for other non-coding regions (Zhang et al. 2007).

The T_s/T_v ratio in numts from the MCR was significantly higher than numts from other mitochondrial regions. As numts are non-functional and randomly distributed in all the chromosomes, the detected differences in substitution patterns between numts are most likely due to the nucleotide composition of each group (Leister 2005; Gherman et al. 2007). Genomic regions rich in GC usually exhibit elevated T_s/T_v ratios and (Bulmer et al. 1986; Hess et al. 1994), as shown here, GC content of the MCR is larger than the mitochondrial average, which could explain the greater T_s/T_v ratio in numts derived from the MCR relative to other numts.

The comparison between PVS and numt abundance in humans allowed us to conclude that mitochondrial genes with relatively conserved sequences and probably under stronger stabilizing selection may maintain their identity with nuclear copies much more than those evolving under a model of drift. As a result, these sequences may be much less divergent from their nuclear copies and consequently appear to be more prevalent in the nuclear genome.

The presence of one or more microhomologies between individual numt termini and their corresponding integration site was observed in nearly half of all studied mitochondrial - nuclear junctions, suggesting that NHEJR is a predominant mechanism of numt integration. This result reaffirms previous studies suggesting that mitochondrial fragments frequently insert into sites with double strand breaks, which are then subsequently ligated through NHEJR (Blanchard & Schmidt 1996; Ricchetti et al. 1999; Hazkani-Covo and Covo 2008; Jensen-Seaman et al. 2009). Microhomology abundance might actually be underestimated since substitutions in the numt boundaries or in the mitochondrial sequence will tend to blur the composition of both mitochondrial and nuclear sequences at the time of integration. Although microhomologies are prevalent, the present study found no evidence for identical nucleotide motifs in both numt endings, indicating that *trans*-replication slippage is likely to be unimportant as a mechanism of integration. Previous work has found two instances where this mechanism may have played a role in numt integration (Chen et al. 2005). However, instances of *trans*-replication slippage are likely to be rare as it requires the co-existence of a very specific set of conditions (see Chen et al. 2005).

Microhomologies were also observed between mtDNA fragments tandemly located in certain numt loci and inserted during the same inter-nodal period. The insertion of multiple fragments within the same time period suggests that insertion of such fragments likely happened as part of a single event. Furthermore, the presence of microhomologies between adjacent mitochondrial fragments indicates that their assembly in the nuclear genome is not random but it is likely guided by specific recognition between fragments through two potential mechanisms: non-homologous recombination (Farrelly & Butow 1983) or NHEJR of isolated fragments imported into the nucleus during episodes of occasional mtDNA “leakage” or intensive mitochondrial degradation (Kamimura et al. 1989).

Our study is the first to test for the temporal concordance in insertions of mitochondrial fragments and TEs. In order to identify cases where a TE-LINE1 transduced a numt, both elements need to have the same approximate age and be located adjacent to one another. Individual analysis of over 80 great ape numts failed to find any signature of TE-LINE1 3' transduction in the duplication of numts, strongly suggesting that this mechanism is unimportant in the insertion of numts in Hominoidea. In a few cases, TEs were found to match the age of a flanking numt but these events were very rare. Thus, these observed TE-numt matches may not reflect any real association since the prevalence of TE families surrounding the insertion site of numts in the human genome reflects their prevalence across the whole genome.

In summary, this study provided valuable information on three important aspects of numt molecular evolutionary dynamics which we summarize here: 1) An excessive T_s/T_v ratio in numts derived from non-coding mitochondrial regions of the mitochondrial genome, i.e. MCR, that could be attributable to the relatively large GC content of this mitochondrial region. This suggests that the large GC content in MCR numts could affect patterns of nucleotide substitution and favor an increase in the number of transitions over transversions. However, further studies should address this issue in more detail. 2) An observed excess in the number of numts from mitochondrial genes under strong stabilizing selection and an apparent deficit of numts from the non-coding MCR (Soto-Calderón et al. *in review*) shows that comparisons of rates of insertion of fragments originating from different parts of the mitochondrial genome should take into account the rates of nucleotide substitution in the mitochondrial genome. 3) NHEJR is the most predominant mechanism of integration across multiple taxa and our ability to date times of insertion allows us to definitively rule out the importance of TEs and *trans*-replication slippage in expanding the numt population in great apes.

CHAPTER 3

ISOLATION OF GORILLA-SPECIFIC NUCLEAR INSERTIONS OF MITOCHONDRIAL DNA (NUMTS) AND THEIR POTENTIAL AS POPULATION GENETIC MARKERS.

Abstract

Although the mitochondrial control region (MCR), and specifically the First Hyper-Variable domain (HVI), has been a widely used as molecular tool in population genetics, rampant amplification of nuclear translocated copies (numts) in gorillas has compromised the reliability of mitochondrial sequence databases. Previous studies of MCR variation in gorillas indicate that all putative MCR HV1 numts fall into three distinct classes (I, II and III) which appear to be entirely gorilla-specific. However, the identity, number and location of these loci in the gorilla genome is completely unknown, thus preventing the systematic study of numt diversity and design of locus-specific primers. In order to address these questions, we conducted BLAST searches of the gorilla genome by using the whole mitochondrial genome as query sequence and by screening two types of genomic libraries with HVI MCR numts. Five gorilla-specific numts were isolated and mapped. Four of these loci contain HVI (Numt 1_1, Numt 2_1, Gcl18_1 and CABD5746) and one (Go11_188) contains other MCR sub-domains and pseudogenes flanking this domain. Both Numt 1_1, Numt 2_1 contain the entire HVI and showed high similarity with numt classes IIb and I, respectively. Amplification of all five loci from captive zoo animals with locus-specific primers allowed the identification of insertional polymorphisms for three of them (Numt 1_1, Numt 2_1, Gcl18_1). Preliminary data also indicate their potential utility as nuclear molecular markers for future tests of phylogeographic models inferred from mitochondrial markers and morphological data.

Introduction

For decades, studies of population genetics and systematics have relied heavily on mitochondrial DNA (mtDNA). However, the unintentional amplification of nuclear copies of mtDNA (numts) can mislead mitochondrial analyses either through the overestimation of genetic diversity (see Garner & Ryder 1996) or through erroneous phylogenetic inference (Song et al. 2008). Assembly of comprehensive animal genomic databases has shown that the number of numts varies among taxa (Hazkani-Covo 2010). Although numts are rare in some animal taxa such as *Anopheles* mosquitos, chickens and rats (Richly & Leister 2004), they appear to be widespread in many primates including African great apes (Triant & DeWoody 2007).

Nowhere is the problem of numt contamination of mitochondrial databases more acute than in gorillas (Jensen-Seaman et al. 2004; Calvignac et al. 2011), leading some to question the reliability of mitochondrial sequences in this taxon (Thalmann et al. 2004; 2005). This problem is all the more complex because of the apparently high incidence of *in vitro* recombinants which have also been misdiagnosed in earlier studies (Anthony et al. 2007a). As most primate numts are usually <500b, it may be possible to avoid their inadvertent amplification in regular polymerase chain reaction (PCR) amplifications through the use of long-range PCR (Thalmann et al. 2004). However, this approach is not always feasible as degraded samples from feces or museum specimens are frequently the only source of DNA for gorilla genetic studies.

Previous phylogeographic studies of the first hyper-variable (HV1) domain of the mitochondrial control region (MCR) in gorillas have identified four major mitochondrial haplogroups (A - D) and three different numt classes (I - III) that overlap this region (Anthony et al. 2007a; 2007b). Whereas mitochondrial haplogroups A and B are restricted to east mountain (*G. beringei beringei*) and east lowland (*G. beringei graueri*) gorillas, haplogroups C and D are only found in western gorillas (*G. gorilla*) and are also restricted in their geographical distribution: Haplogroup C is largely limited to Nigeria and Cameroon whereas haplogroup D is found in Gabon, Equatorial Guinea, The Republic of the Congo and the southern tip of the Central African Republic (CAR) (Gagneux et al. 1999; Grubb et al. 2003; Clifford et al. 2004; Anthony et al. 2007a). With respect to numt classes I - III, group I has to date only been found in western gorillas whereas class II and class III numts are present in both east lowland and western gorillas (Thalmann et al. 2004). All of these MCR numts appear to have inserted very recently, making them difficult to distinguish from authentic mitochondrial sequences. Therefore, a definitive characterization of these numt loci and their distribution among major geographic haplogroups in gorillas requires mapping their location in the gorilla genome and sequencing their nuclear flanks (Zischler et al. 1995b; Thomas et al. 1996).

Identification of locus-specific primers will also allow the specific amplification and analysis of insertional polymorphisms of recent numts. A survey of the pattern of these insertional polymorphisms in samples of different geographic origin will also provide important information on their potential as population genetic markers. In contrast to other polymorphic markers (e.g. STR's, AFLPs, SNPs), loci with insertional polymorphisms such as transposable elements (TEs) and numts are considered homoplasy-free markers because they are rarely excised from the genome and allow the ancestral (absence) and derived (presence) states to be inferred (Batzer & Deininger 2002). In the past, insertional polymorphisms in Transposable Elements have been intensively used to study historical demography of human populations (Perna et al. 1992; Batzer et al. 1994; Batzer & Deininger 2002; Herke et al. 2007). With the advent of whole genome sequencing, numerous polymorphic numts have now been identified in humans (Thomas et al. 1996; Lang et al. 2011). These markers have proved useful in assessing previous phylogeographic hypotheses based on mtDNA and as a means of assessing levels of genetic admixture between populations. Despite these advances in human genetics, few studies if any have addressed the utility of numts as population genetic markers in gorillas and non-human primates. By adopting a systematic survey of numt loci either through available genomic resources or more conventional library based approaches, it is now possible to identify and characterize numt loci in gorillas and in so doing (i) conduct quality control of HVI databases (ii) identify potentially polymorphic loci that can be used as nuclear markers and (iii) complement previous phylogeographic studies based on mtDNA (Jensen-Seaman et al. 2001; Anthony et al. 2007a; 2007b).

To do this, we carried out genomic BLAST searches of the current gorilla scaffold and combined this with targeted screens from a BAC library. As both methods are only based on one individual, we also made use of an anchored PCR survey of genomic DNA enriched from multiple unrelated individuals in order to capture any additional loci that may not have been present in the initial screens. The main goals of this study were therefore to: (1) isolate and map all gorilla-specific numts present in the public database of the gorilla genome; (2) screen a commercial BAC library and survey nuclear enriched genomic DNA obtained from 5 individual gorillas through an anchored PCR assay (A-PCR) in order to isolate additional MCR numts; (3) examine the prevalence of these numts in previous published databases (Garner & Ryder 1996; LaCoste et al. 2001; Jensen-Seaman et al. 2004; Thalmann et al. 2004; 2005); (4) assess patterns of insertional polymorphism in a sample set of captive gorillas of known mitochondrial haplogroup association and; (5) provide recommendations on how such markers could be employed in future studies of gorilla phylogeography.

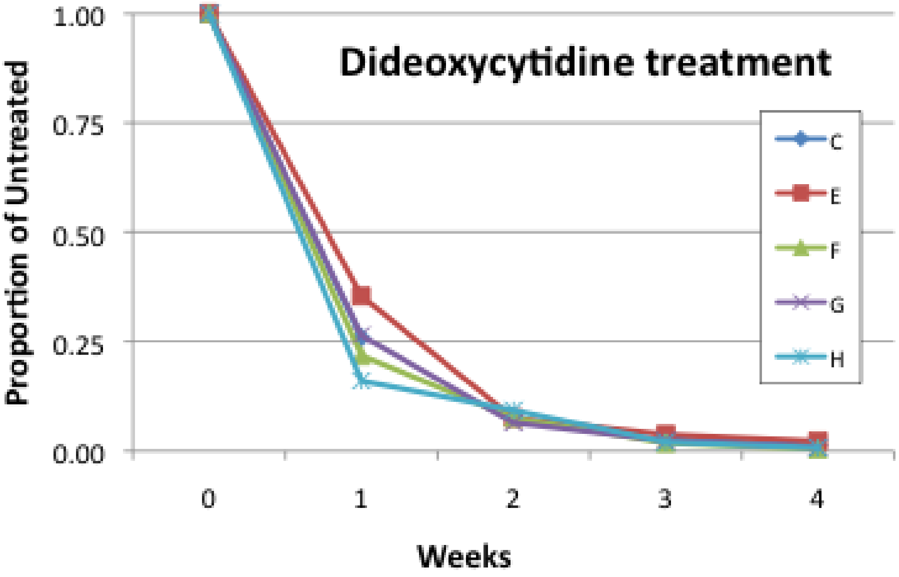
Materials and Methods

BLAST search. The entire gorilla mitochondrial genome (X93347) was used as a query sequence in BLASTn searches (Altschul et al. 1990) of the partially assembled reference genome (*Gorilla gorilla gorilla*, CABD00000000.2, November 2009) and the original trace files (Gorilla_gorilla_WGS) of the western gorilla Kamilah. Only hits of either i) at least 100bp in length and 60% identity or ii) a size of between 50 and 99bp and an identity greater than 70% were considered for further analyses. Flanking sequences of positive numt hits were aligned with the reference genome of humans (build 36.3; 2006) and chimpanzees (build 2.1; 2006) available through the UCSC website (http://genome.ucsc.edu/cgi-bin/hgGateway), to map their genomic location and determine whether these insertions were unique to the gorilla (i.e. absence in humans and chimpanzees).

BAC library screens. Nine different radiolabeled probes designed from the gorilla MCR and flanking tRNA sequences were hybridized to a genomic BAC library derived from the western gorilla Frank (library CH255, obtained from the Children's Hospital Oakland Research Institute, Oakland, CA). Thirty-two positive BACs were subsequently grown up and characterized in an effort to isolate gorilla-specific numts. The BAC-ends were sequenced and mapped to the human genome sequence. Those BACs that overlapped with a known human numt were not further characterized. Those that did not overlap were deemed likely to contain a gorilla-specific numt and were further characterized with internal sequencing, primer walking, and/or sub-cloning in order to determine the complete numt sequence along with the flanking unique nuclear sequence.

Isolation of nuclear-enriched gorilla DNA and Anchored PCR. Several complementary strategies were adopted in order to bias amplification of nuclear copies relative to mtDNA. Firstly, we took advantage of the rarity of *Bgl*II (A↓GATC↑T) recognition sites in the gorilla mitochondrial genome to favor amplification of nuclear targets. This is because linker ligation and subsequent PCR amplification of target DNA during A-PCR (see below) requires restriction enzyme digestion. We surveyed for the presence of *Bgl*II sites in the entire mitochondrial genome of eight different gorilla fibroblast cell lines (Coriell Institute for Medical Research, Camden, NJ) by amplifying the entire mitochondrial genome in three overlapping fragments through long-range PCR with the enzyme *LA Taq* polymerase (TaKaRa Bio Inc., Mountain View, CA). Primers employed in this step are listed in Table 3.1 a. PCR products were then digested with *Bgl*II (NEB) and five individuals that lacked *Bgl*II restriction sites were then used for the following steps of nuclear DNA (nDNA) enrichment (see a list of selected gorillas in Figure 3.1).

Figure 3.1. Reduction in the ratio of mtDNA to nDNA in gorilla fibroblasts grown in the presence of 2',3'-dideoxycytidine(ddC). The ratio of mitochondrial to nuclear amplification products is shown relative to untreated controls for each time point for each of five gorilla fibroblast cell lines (C: Chipua [PR00622]; E: Billy [PR00671]; F: JimmyJr [PR00943]; G: Chaka [PR01013], H: Kimya [PR01023]).



The five selected cell lines were then treated with 2',3'-dideoxycytidine (ddC), which impairs mtDNA replication resulting in a progressive dilution and virtual loss of mtDNA in cultured cells (Ashley et al. 2005). To do this, each cell line was cultured in fibroblast culture medium containing 10% fetal bovine serum (FBS), 0.9X Eagle's minimal essential medium (EMEM), supplemented with 10 μM ddC and 205 μM Uridine. Cell cultures were grown at log phase at 37˚C in a humidified 5% CO_2 incubator changing the culture medium regularly. Treatment continued for four weeks at which time DNA was extracted from cell lines using the Blood & Cell Culture DNA Maxi Kit (Qiagen, Valencia, CA). This kit recovers high-molecular-weight DNA with an average length of 50-100 Kbp, further reducing the abundance of any remaining mtDNA (~16.5 Kbp). The ddC treatment and preferential extraction of nDNA led to a reduction of 90% in the relative copy number of mtDNA vs nDNA in treated cells as compared with untreated cells (Figure 3.1). The copy number of mtDNA relative to nDNA was estimated by quantitative PCR of two reporter genes, the mitochondrial cytochrome *b* gene and the nuclear tumor suppressor gene p53, amplified with the specific primer pairs CytbGor F / CytbGor R and p53iiPrim F / p53ii-R respectively, using the SYBR® Green PCR Master Mix in a ABI StepOnePlus™ Instrument (Table 3.1 a).

Table 3.1 a. List of primers used to: (i) Amplify the complete mitochondrial genome through long-range PCR of three overlapped segments; (ii) Estimate the relative copy number of mitochondrial vs. nuclear DNA (qPCR); (iii) amplify gorilla-specific numts (Numts) and an internal PCR standard (PCR control); (iv) sequence the HVI subdomain.

Primer F		Primer R	
(i) Mitochondrial long-range PCR:			
mt10261 Fa	atcaacacaaccacccacagccta	mt726 R	ggctacaccttgacctaacgtctt
mt551 F	actgctcgccagaacactacgagc	mt7969 R	ggtaagcctaggattgtgggggca
mt7022 F	tgcagcgcaagtaggtctacaaga	mt12400 R	gctgatttgcctgctgctgctagg
(ii) qPCR:			
CytbGor F	taacggcgcctcaatattct	CytbGor R	gtaggaggatgatgccgatg
p53iiPrim F	ggagcactaagcgaggtaagc	p53ii R	ggaaagaggcaaggaaaggt
*(iiia) Numts: **			
Numt1_1 Fa	attacagacgcacgccacca	Numt1_1 Ri	tagcattgcgaaacgctggaacc
Numt2_1 F2	tgatgcccctcctccaatctgtg	Numt 2_1 Ri	tttcgacgggctcacatcaccc
Numt2_1 F1	ccagtcattgagcatgtacttccct	Numt2_1 R2	ttggggcaaatattggtctctg
Gcl18_1 F	gatctctcttcttttccattggtc	Gcl18-1 R	gaggcattccattacccaac
Gcl18_1 Fi	cgacctgcctcctacaaaag	Gcl18-1 R	
CABD5746 Fi	cgattgctgtacgtgcttgt	CABD5746 R2	cagtttgggtttggtttgct
Go11_188 F	catgctcttatgggcctgaa	Go11_188 Ri	cggcatctggttcttacttgag
Go11_188 Fi	cagatgccggatacagttcatt	Go11_188 R	cctctgattctcttgcaggttg
(iiib) PCR Control:			
TP53 F	aagggtggttgggagtaga	P53ii R	ggaaagaggcaaggaaaggt
p53 3F	cactggaagactccaggtcag	P53ii R	
(iv) HVI sequencing primers:			
mt15365 F	ccttccaagggcatattcag	mt15888 R	ttaaggggaacgtgtgaagc

(*) Numt primers were combined with the primers TP53 F / P53ii R, which were used to amplify an internal standard. The only exception was the case of Numt 2_1 F1 / Numt 2_1 R2, where the internal standard was amplified using the primers P53 3F and P53ii R.

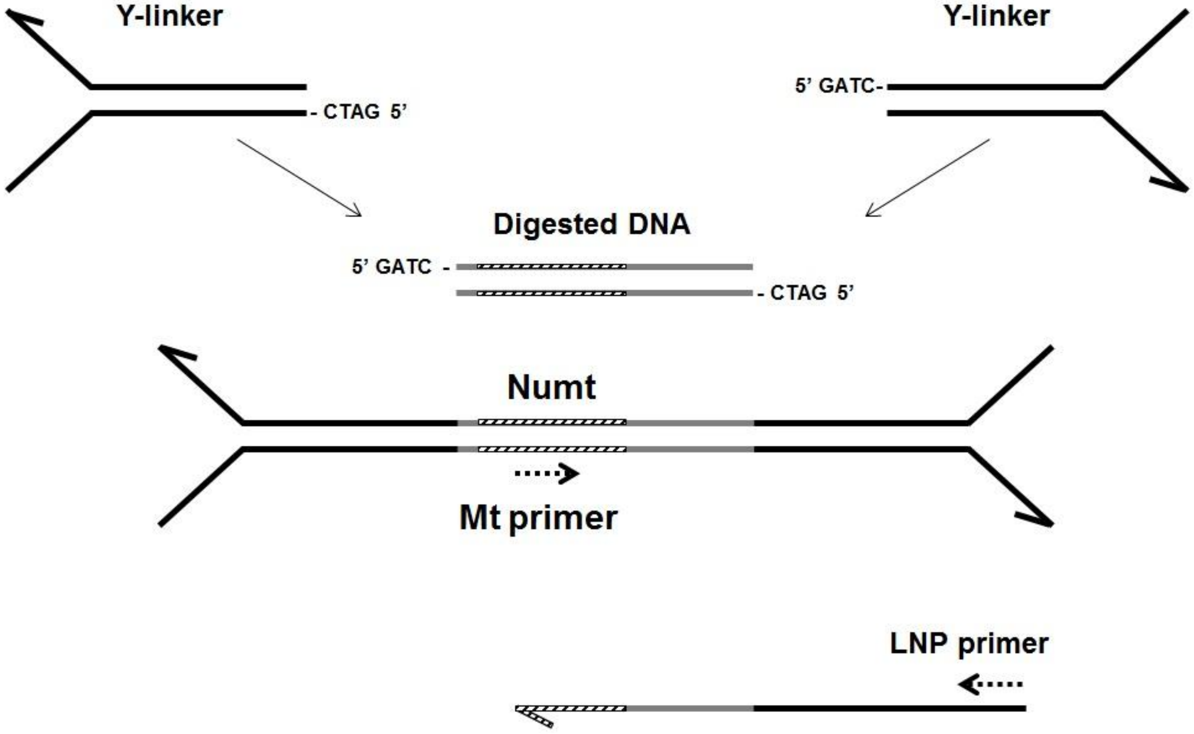
Table 3.1 b. The two parts of the Y-linker and the LNP primer used in the anchored PCR (A-PCR).

A-PCR oligos:	
*Bgl*II-top	gatcgaaggagaggacgctgtctgtcgaagg
Bottom	gagcgaattcgtcaacatagcatttctgtcctctccttc
LNP	gaattcgtcaacatagcatttct

The nuclear-enriched DNA samples obtained from the five treated cell lines were subsequently pooled in equivalent proportions and completely digested with *Bgl*II to generate fragments with 5'-GATC overhangs. Digested DNA was then ligated to a compatible Y-linker made of two partially complementary oligos (*Bgl*II-top and bottom; Table 3.1 b), modified from Ray et al. (2005). Fragments containing the MCR were selectively amplified through an A-PCR assay using one of several MCR primers in combination with the LNP primer (Figure 3.2; see Ray et al. 2005 for details). During the first PCR round, the MCR primer binds to a given fragment containing a numt leading to the extension of the first strand (Figure 3.2). This creates a binding site for the LNP primer at the 3' end of this new strand, allowing the amplification of the complementary reverse strand. This step is followed by a semi-nested PCR using the same LNP primer and an internal MCR primer to increase the specificity of the PCR amplification.

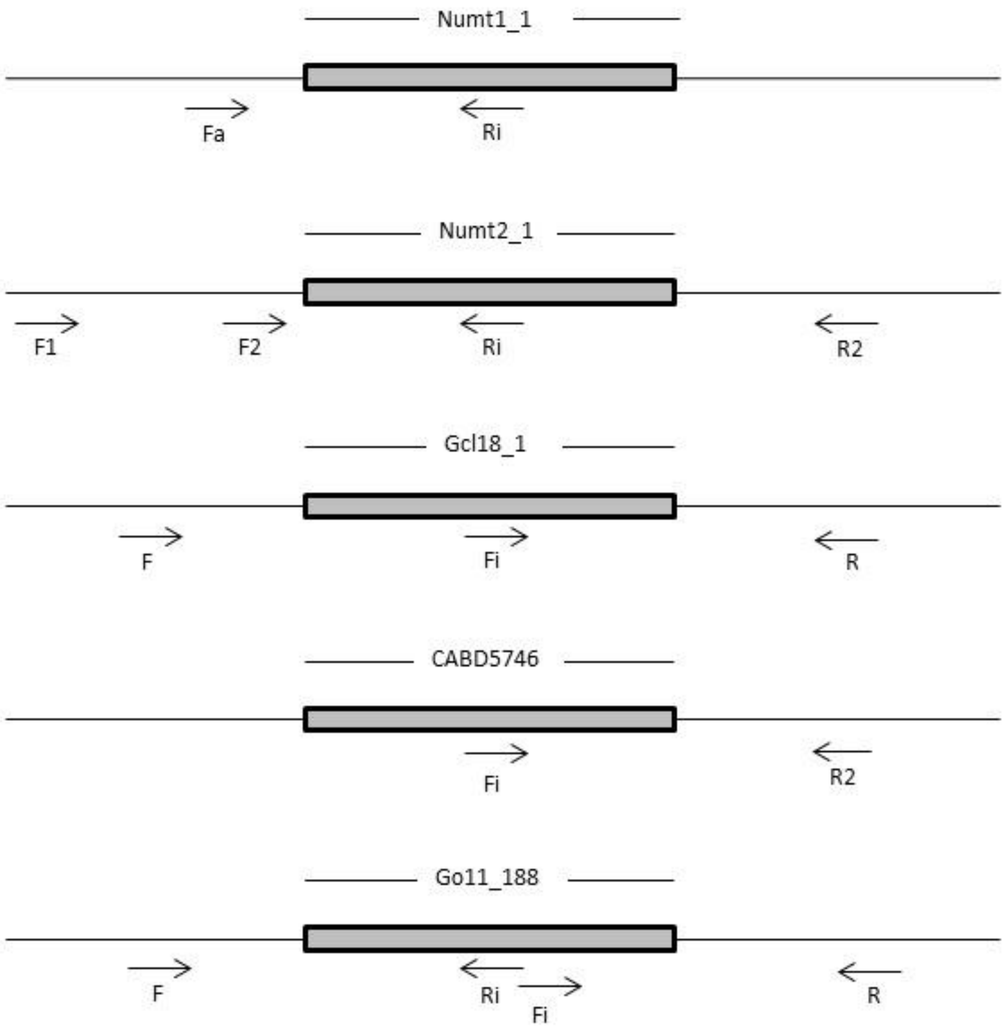
A-PCR amplifications were carried out in 20μl reactions containing 1U *LA Taq* polymerase, 0.4mM dNTPs, 0.2μM of each primer, 1X buffer and 20-30ng DNA. Cycling consisted of initial denaturation at 94°C, followed by 35 cycles of 94°C for 15 s and 68°C for 15 min, with a final extension at 72°C for 2 min. PCR products were then cloned into the pCR®2.1 vector using the TOPO TA-cloning kit (Invitrogen) and sequenced with the Big-Dye v1.1 (ABI). These sequences were then aligned with the gorilla mitochondrial genome to determine the extent of the 5' portion of the numt and its adjacent nuclear flank. Numt sequences were then BLATed (Kent 2002) against the human and chimpanzee reference genomes to identify the genomic location of the orthologous locus, infer the sequence of the second flank and determine whether the corresponding numt was unique to gorillas. Lastly, primers flanking the numt were designed to amplify the remaining part of the numt.

Figure 3.2. Representation of the anchored PCR protocol used to select and amplify genomic regions containing numts. Compatible 5' overhangs are shown in both the Y-linker (solid black lines) and digested DNA fragments containing a numt (dotted lines). Recognition sites of mitochondrial LNP primers and direction of DNA replication are shown with dotted arrows.



Identity of mapped HVI numts and previous numt reports. In order to determine the identity of the gorilla MCR numts obtained in the study, the portion of these loci that encompassed the first hyper-variable region (HVI) was aligned with a set of 41 previously reported gorilla HV1 MCR numts (Classes I, II and III) (Anthony et al. 2007a). Sequences were aligned using the Clustal W program as implemented in MEGA v5 (Tamura et al. 2005). A poly-cytosine stretch of 26bp that is prone to error during polymerase amplification was deleted from the sequence alignment prior to phylogenetic analysis. A stretch of 90bp unique to the mapped numt CABD5746 was also removed from the alignment. The program jModelTest v.0.1.1 selected TPM3uf+G model as the most likely model of nucleotide substitution (Posada 2008). This model of substitution was used to construct a Neighbor-Joining tree in PAUP* v.4.0b10 (Swofford 2002), using an alpha shape parameter value set to 0.375 and no invariable sites. Bootstrap support for individual branches was estimated with 1000 replicates, retaining branches with 50% support or greater.

Figure3.3. Relative location and orientation of the primers (arrows) used to amplify gorilla-specific numts (gray boxes).



Analysis of insertional polymorphisms. Once flanking sequences were obtained for all gorilla MCR numts, their presence/absence was assessed using a panel of 68 DNA samples donated by US zoos and collaborators (Appendices F and G). All genomic DNA was extracted from peripheral blood using the DNeasy Blood & Tissue Kit (Qiagen). Where possible, primers were designed to amplify the entire region containing the numt. This approach allows the discrimination of individual genotypes whose alleles can be separated by size i.e. a larger product indicating a numt insertion (+) in one or both alleles or a smaller product (-) indicating the absence of the insertion. Amplification of the entire region containing a numt was not always possible for all loci due to the large size of the target region or potential primer disparities with the annealing site. In such cases primers were designed to amplify a portion of the numt and one of its flanks (Figure 3.3). In these cases, only the presence or absence of the numt was then possible (see Numt1_1 in the Results section). To safeguard against false negatives (i.e. failure of the PCR reaction), all PCR reactions were carried out using an internal standard based on co-amplification of a conserved region in the housekeeping tumor-suppressor gene p53 (Table 3.1 a). A human sample was also amplified in each experiment as a negative control. Each 20μl PCR reaction contained 0.5U of *Taq* DNA polymerase, 2.5mM $MgCl_2$, 200μM dNTPs, 250μM each primer and 20-30ng DNA and 1X PCR buffer (Invitrogen). Cycling consisted of 2 min of initial denaturation at 94°C followed by 35 cycles of 94°C for 30s, 58-64°C for 30s and 72°C for 50s - 2 min, with a final extension at 72°C for 2 min. PCR products were run in 2-4% agarose gels.

Table 3.2. Description of the five mapped gorilla-specific numts.

Numt Name	Isolation method	Numt Size	Mt position (X93347)	Sample Size	Polymorphic Status
Numt1_1	BAC screening BLAST - Trace files	1400	14806-16150	62	Yes
Numt2_1	BAC screening BLAST - Trace files	2500	141-16412; 1-14060	65	Yes
Gcl18_1	A-PCR	450	15530-15993	58	Yes
CABD5746	BLAST - Contigs	450 *	15615-15904; 8430-8466;	10	Fixed
Go11_188	BLAST - Trace files	2350	15788-16412; 1-1720	14	Fixed

(*) This includes a fragment of 90bp lost from the gorilla mitochondrial genome.

As the mitochondrial haplogroups of gorillas exhibit geographical affiliation, the mitochondrial lineage of wild gorillas was used to discern potential geographical differences in the natural distribution of polymorphic numts. To do this, we identified 17 wild-born captive gorillas and one whose ancestors possessed the same mitochondrial lineage. The HVI haplogroup of these gorillas or a relative with the same maternal lineage (Wharton 2007) was established by mitochondrial sequencing. In order to obtain these HVI sequences, we first long-range-PCR amplified 6,880bp of the mitochondrial region containing the HVI sub-domain using specific primers (mt10261 Fa / mt726 R; Table 3.1). Both strands of the PCR products were then sequenced using primers flanking the HVI sub-domain (mt15365 F / mt15888 R) and Big-Dye kit v1.1 (ABI). Sequences were then combined with a reference HVI sequence database (n=166) from Anthony et al. (2007b) and an additional sequence dataset of free-range gorillas and captive gorillas generated for the present study. A total of 231 gorilla HVI sequences were aligned in MEGA5 (Tamura et al. 2011) and the nucleotide substitution model was selected with jModelTest v.0.1.1 (Posada 2008). The program jModelTest v.0.1.1 selected TPM1uf+G which was used with an alpha shape parameter of 0.360 to construct a Neighbor-Joining tree in PAUP* v.4.0b10 (Swofford 2002). Only nodes with bootstrap support values of 50% (500 out of 1000 repetitions) or greater were retained. Clustering of the target captive gorillas with the traditional mitochondrial haplogroups allowed their assignment to a specific mitochondrial lineage.

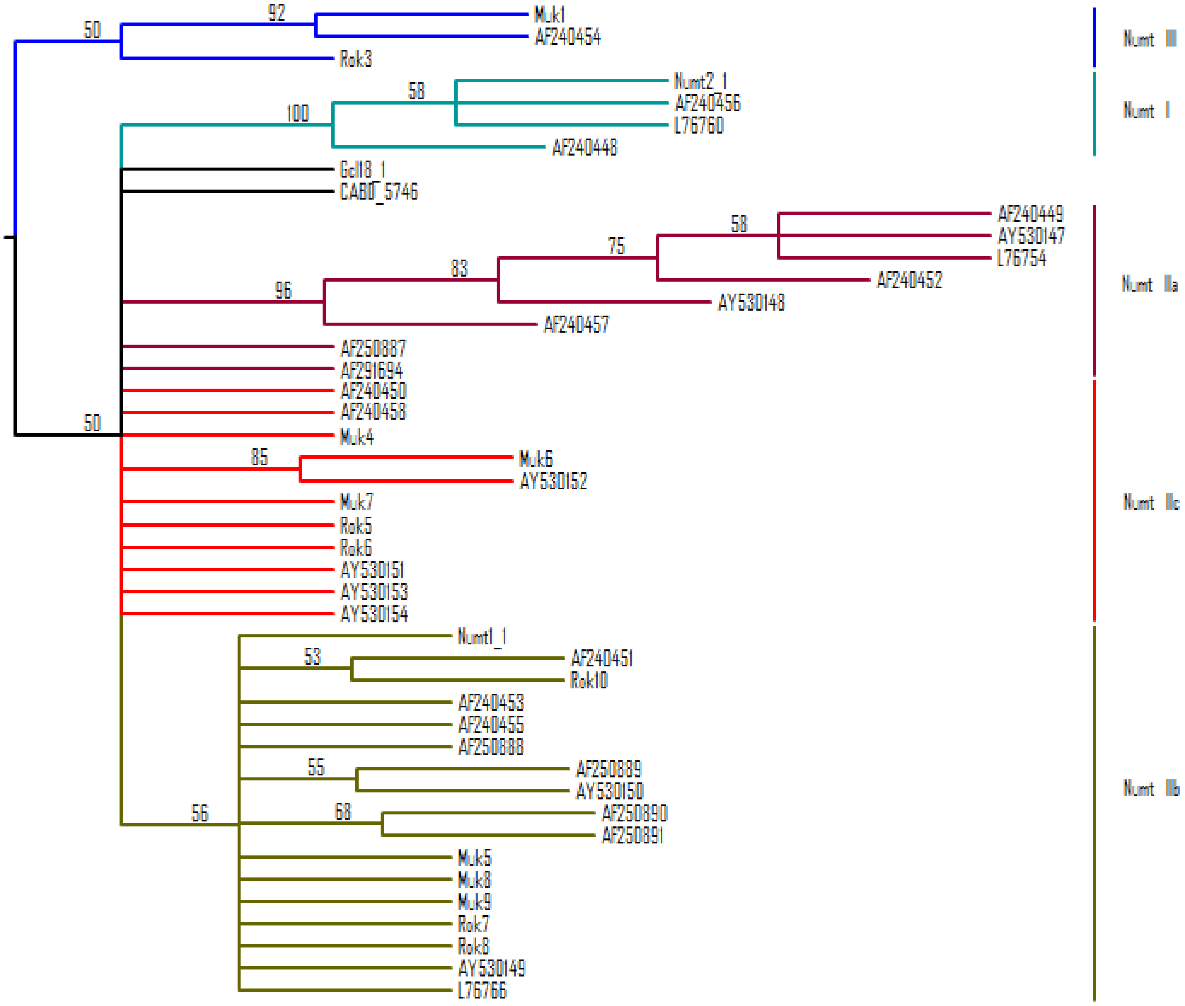
Results

A total of 22 putative gorilla-specific numts were found with one or more searching strategies (Table 3.2; Appendix H). Seventeen of these putative gorilla numts either failed to amplify or appeared to be absent in the captive gorilla sample set available (Table 3.2). The five remaining numts successfully amplified from the gorilla genomic DNA panel. These comprised four loci containing the HVI sub-domain (Numt 1_1, Numt 2_1, Gcl18_1 and CABD5746) and one (Go11_188) that contained other MCR sub-domains as well as additional mitochondrial sequence (gorilla mitochondrial genome X93347: 15788-16412; 1-1720). Two loci (Numt 1_1 and Numt 2_1) were found in both the BAC library screens and BLAST searches of the original short reads (trace files) of the gorilla genome. Gcl18_1 was identified through an A-PCR approach whereas CABD5746 and Go11_188 were only found through BLAST searches of gorilla contigs and trace files, respectively (see Table 3.2 for details).

Numt1_1 is an insertion of ~1,400bp that contains most of the MCR including HVI, the Central Conserved Domain (CCD) and an extensive portion of the second hyper-variable region (HV2). Neighbor Joining analysis of HVI numts showed that Numt1_1 was very similar, if not identical to representatives of Class IIb numts found in both western and eastern lowland gorillas (Anthony et al. 2007b) (Figure 3.4; Appendix I). Sequences showing the highest identity with this numt include AY530149, isolated from a wild gorilla in Lobéké, Cameroon (Clifford et al. 2004); L76766, from the captive gorilla Carolyn, captured in the Congo region (Garner & Ryder 1996) and Rok8 from the captive gorilla Rok (Thalmann et al. 2004), which with the exception of a 4bp gap showed perfect identity with Numt1_1. Despite the great similarity of these two numts, our data showed that Numt1_1 is not present in the gorilla Rok, indicating that

Rok8 and Numt1_1 may represent different alleles. Perfect identity with Numt1_1 was also observed with two numts found in two different eastern lowland gorillas including AF240455 (LUT2DTA9) (Lutunguru, Democratic Republic of Congo; Jensen-Seaman et al. 2004) and Muk5 (Thalmann et al. 2004).

Figure3.4. Neighbor-Joining tree of a sequence database showing the relationship of mapped HVI numts (Numt1_1, Numt2_1, Gcl18_1 and Numt5746) and 41 reference numt sequences in Classes (I – III), as defined by Anthony et al. (2007b). This phylogeny was built using a TPM3uf+G model (alpha = 0. 0.375). Numbers indicate bootstrap support values (≥50%).



Numt2_1 is an insertion of ~2,500bp that contains the entire MCR region. The HVI portion of this numt sequence clustered with Class I numts and exhibited sequence identities of 99% with the eastern lowland gorilla sequences AF240456 (LUT2DTA10) and AF240448

(LUT2DTA1) (Appendix I) (Jensen-Seaman et al. 2004). This numt also exhibited high sequence identity (97%) with L76760, from the western gorilla Jojo (Garner & Ryder 1996). Curiously, Numt2_1 also shared elevated sequence identity (99%) with a western lowland gorilla sequence AY530145 (BEL1a), considered to be an *in vitro* recombinant between mitochondrial and class I nuclear templates. Both of class I numts and western gorilla sequences are very similar in composition (see Anthony et al. 2007b), making them particularly easy to misidentify.

The numt Gcl18_1 is a ~460bp insertion that encompasses the entire CCD along with a section of both HVI and HV2 (Appendix I). It does not show close resemblance with any other HVI numt previously described. The HVI portion of this numt shared sequence identity as high as 95% with the class IIc numts Muk4, Muk6 and Muk7, amplified from a single eastern lowland gorilla (Muk), and Rok5 amplified from the western gorilla Rok (Thalmann et al. 2004).

The numt CABD5746 consists of an insertion of 250bp made up of two non-contiguous mitochondrial fragments, including one containing a portion of HVI and the CCD and another one 7,450bp apart in the mitochondrial genome. Although CABD57646 is exclusive to gorillas, it contains a 90bp section of mtDNA that is no longer present in the mitochondrial genome of contemporary gorillas but still found in all the other great ape taxa. Like Gcl18_1, CABD57646 could not be assigned to any predefined numt class and it is substantially different from previously reported HVI numts (<85% identity).

Finally, the longest gorilla numt found in this study is Go11_188. This mitochondrial insertion of ~2,350bp contains all MCR subdomains other than HVI as well as copies of the mitochondrial 12S and 16S rRNA genes and the phenylanine and valine tRNAs.

A detailed comparison between Numt1_1 and Numt2_1 sequences and their corresponding mitochondrial sequences was also carried out in order to assess the extent to which existing mitochondrial primers would be expected to co-amplify the corresponding nuclear copy (see Figure 3.5). The following primer pairs have been previously used in earlier studies: MirRev4/ProFor2 and D-441/D88 (Jensen-Seaman et al. 2004); H402/L91 (Garner & Ryder 1996); H16498/L15926 (Thalmann et al. 2004); and MTD1AS/MTD1S (LaCoste et al. 2001). We also note that H16498 is the same primer as MTD1AS. The nuclear regions that align with these primers are virtually identical in the mitochondrial and nuclear copies and in no case favor amplification of the mitochondrial over the nuclear copy. In four cases, the primer sequences were identical to both the mitochondrial and the nuclear copies (H16498, MidRev4, D-441 and ProFor2). In two cases there was only one internal mismatch between the primer and mitochondrial/nuclear copies (H402 and MTD1S). In the remaining three primers, mismatches with mitochondrial and nuclear target sequences were observed in three (D-88), four (L91) and seven (L15926) nucleotide positions. However, in the latter case all mismatches were concentrated in the 5' region that is not as critical as the 3' in terms of primer specificity.

Figure 3.5. Alignment of 5'-3' sequences from mtDNA (X93347.1), gorilla numts containing the entire HVI region (numt1_1 and numt2_1) and primers used by previous authors to amplify the mitochondrial copy of this region in gorillas.

```
X93347.1         CCTGAAGTAGGAACCAGATGCCGGATACAGT
Numt1_1          ....................T..........
Numt2_1          ...............................
H16498/MTD1AS    ....................
MidRev4                 ........................

X93347.1         CGGGATATTGATTTCACGGAGGATGGTGTTC
Numt1_1          ...............................
Numt2_1          ...............................
D-441            ......................
H402                   .....A...................

X93347.1         GTCTCCCCATGAAAGAACAGA-GAATAGT
Numt1_1          .....................-.......
Numt2_1          .....................A.......
D-88 *           .CT..................-.
L91 *            ...................A.-.CT....

X93347.1         GGTGGAGTCGAGGACTTTTTCTCTG
Numt1_1          .........................
Numt2_1          .........................
ProFor2 *        .........................

X93347.1         AGCTTTGGGTGCTGATGGTGGAGTCGAGGACTTTTTCTCTG
Numt1_1          .........................................
Numt2_1          .........................................
MTD1S *          .............A......
L15926 *                       ...................AGCT.TGA
```

(*) Reverse primer sequence.

From the five gorilla numts mapped in this study, only Numt1_1, Numt2_1 and Gcl18_1 were found to be polymorphic in captive western gorillas. Both CABD5746 and Go11_188 were always found to be present and are probably fixed, at least in western gorillas (see Table 3.2 and Appendix G). A total of 18 captive gorilla DNA samples were found to either have been derived from parents of the same haplogroup (n=1) or were known to be wild-born (n=17). These samples possessed haplogroups belonging to four out of five previously defined mitochondrial lineages in haplogroups C (C1 and C2) and D (D2, D3) (see Table 3.3; Figure 3.6). Interestingly, another previously undescribed sub-lineage (C3; Figure 3.6) was also represented in this sample. Despite the limited sample size, some noticeable differences in numt insertional polymorphisms were observed between samples from different mitochondrial haplogroups. For instance, Numt1_1 was only found in two out of 15 individuals in all the mitochondrial haplogroups. Numt2_1 is present in half of the assessed individuals and is mainly associated with the sub-haplogroups C1 and D3, but it is absent in C2 and D2. Finally, Gcl18_1 was found in 13 out of 17 individuals and seems to be the most prevalent of all three numts as it is present in all the gorillas with sub-haplogroups C1, C2 and D2, in three out of four individuals with D3 and is only absent in representatives of C3. Again, this analysis is based on a small sample where the apparent differences in the prevalence of these numts among lineages may be influenced by sampling errors. These results nevertheless highlight the polymorphic nature of these numts and indicate differences in their prevalence among geographic regions.

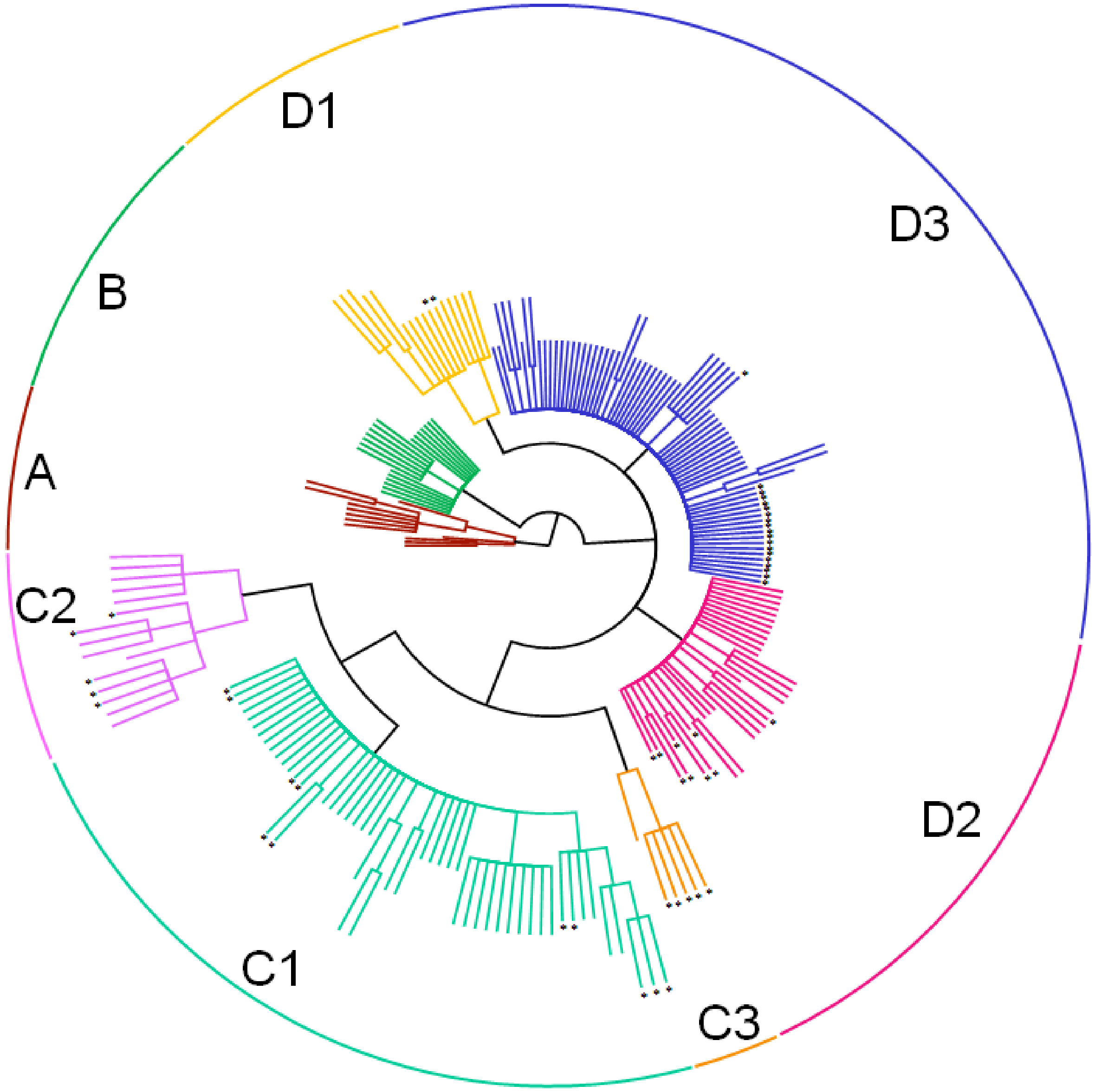
Table 3.3. Proportion of individuals bearing each of the three gorilla-specific numt loci (n=18). Proportion of total assessed chromosomes is shown in parentheses. See Appendix G for a detailed description of gorilla identity and genotypes.

	C1	**C2**	**C3**	**D2**	**D3**	**Total**
Numt1_1	1/3	0/3	1/3	0/2	0/4	2/15
Numt2_1	5/5 (7/10)	0/3 (0/6)	1/3 (1/6)	0/3 (0/6)	3/4 (4/8)	9/18 (12/36)
Gcl18_1	5/5 (6/10)	3/3 (4/6)	0/3 (0/6)	2/2 (3/4)	3/4 (4/8)	13/17 (17/34)

Discussion

We successfully identified and characterized five MCR numts unique to the gorilla genome. Despite the numerous reports of putative gorilla numts from HVI, this is the first study that combines experimental and bioinformatic tools to directly isolate nuclear translocations of the mitochondrial genome in gorillas and in so doing determine the size, region of mitochondrial origin and polymorphic status of these numts. The potential for these methods to effectively retrieve more loci of interest was nonetheless constrained by several factors. Firstly, the relatively low frequency of *Bgl*II restrictions sites in the gorilla genome (~ one site every

Figure 3.6. Neighbor-Joining tree of 231 gorilla HVI sequences comprising a database of 166 sequences (Anthony et al. 2007b) and 65 additional sequences generated in this study, including 51 captive gorillas (asterisks) and 14 free range gorillas from Cameroon. This phylogeny was built using a TPM1uf+G model (alpha = 0.36). All the nodes in this phylogeny exhibit bootstraps support values of over 50%. Mitochondrial haplogroups A and B are restricted to east gorillas (*G. beringei*), whereas haplogroups C and D are only found in western gorillas (*G. gorilla*) (Anthony et al. 2007a).



5,460bp; NEB 2004) may have limited the retrieval of numt fragments via A-PCR. Secondly, gaps in the gorilla current genome data base may have included as yet undetected numts (Scally et al. 2012). Lastly, even though we attempted to survey multiple individuals in our A-

PCR numt search, we may have missed rare numts whose detection might require screening a larger sample of gorillas.

Despite these caveats, previous reports of high HVI numt richness in gorilla sequence databases might be flawed due to errors introduced during amplification and subsequent cloning of PCR products in bacterial vectors. *Taq* polymerase error rate should always be taken into account in determining whether or not two given sequences are significantly different (Williams & Knowlton 2001; Frey & Frey 2004; Thalmann et al. 2004; Song et al. 2008). Such PCR and cloning errors may amplify differences between PCR clones as well as obscure our ability to recover sequences previously obtained from a given sample (e.g. Williams & Knowlton 2001; Jensen-Seaman et al. 2004; Thalmann et al. 2005). This problem may be exacerbated by the use of degraded samples such as feces, shed hair (Clifford et al. 2004) and ancient museum specimens (Jensen-Seaman et al. 2004) as source of genetic material, where poor amplification and low-quality genotyping profiles could lead to poor quality of sequence data. Slight differences between sequences may also stem from allelic variation at the same locus or tandem post-integration duplication, which could explain the high diversity of Numt class II loci so far detected in gorillas. The high similarity between some numt classes and their corresponding mitochondrial copies could also complicate identification of numts (Anthony et al., 2007a), making confirmatory studies in the nuclear genome vital. Furthermore, the high number of near-identical numt sequences within a given individual is also likely to be partly due to a combination of allelic variation or duplicated loci and not necessarily due to multiple independent inserts of mitochondrial fragments (Clifford et al. 2004; Jensen-Seaman et al. 2004; Thalmann et al. 2004). Future progress in numt detection, including sequencing of whole genomes (see Lang et al. 2011) and mapping of candidate numts to one or more reference genomes will permit identification of other putative numts not found in the present study, such as those in numt Classes IIa, IIc and III.

In this study we gathered experimental evidence for at least two classes of gorilla numts that have been shown to be prevalent in other PCR-based studies of gorilla genetic variation (Garner & Ryder 1996; LaCoste et al. 2001; Jensen-Seaman et al. 2004; Thalmann et al. 2004; 2005; Clifford et al. 2004; Anthony et al. 2007a). The high similarity between mitochondrial and nuclear primer binding sites also makes co-amplification of both template types and recombinant types very likely. In addition to these numts, we also identified three additional numts containing a partial segment of the gorilla MCR, two of which (Gcl18_1 and CABD5746) are partially overlapping with the HVI mitochondrial domain. Of these two, only Gcl18_1 indicated high similarity to previously diagnosed numt classes. However, unlike Numt1_1 and Numt2_1, it is unlikely that this locus would be inadvertently amplified with standard HVI primers since it only contains a truncated portion of this mitochondrial region. CABD5746 is made of two non-contiguous mitochondrial fragments, a phenomenon that has been frequently observed in other primate numts (Kamimura et al. 1989; Soto-Calderón et al. *In Preparation*). This numt seems to be a relatively old integration in the gorilla lineage as inferred from its dissimilarity from all the other gorilla HVI numts again making its inadvertent amplification unlikely. Interestingly, it is also probable that this sequence represents an older molecular fossil of a pre-existing mitochondrial haplotype that was subsequently lost during gorilla evolutionary history.

Both the inadvertent amplification of numt loci and generation of mosaic sequences due to template switching during the PCR can also lead to biased estimates of mitochondrial diversity (Song et al. 2008; Chung & Steiper 2008). However, numt identity may only be established with certainty when entire sequences of nuclear integrations are available. As we showed here, differences between mitochondrial and nuclear copies in the regions of primer alignment may be so minor as to allow co-amplification of both templates (Garner & Ryder 1996; LaCoste et al. 2001; Jensen-Seaman et al. 2004; Clifford et al. 2004; Thalmann et al. 2004; Anthony et al. 2007a). As a consequence, specific amplification of gorilla mitochondrial HVI is impracticable using existing sets of primers. Fresh tissue should be used when possible as a source of genetic material in order to take advantage of the high number of mitochondrial copies and favor its amplification over nuclear copies. Although fecal material and museum specimens are frequently the only accessible source of genetic material, mtDNA easily degrades in this kind of sample thus increasing the nuclear-to-mitochondrial ratio and consequently the chance of amplifying nuclear templates with conventional mitochondrial primers (Greenwood & Päävo 1999; Berger et al. 2001; Foran 2006; Soto-Calderón et al. 2009). Under these circumstances, we recommend the use of long-range PCR to bias against numts (Thalmann et al. 2004; Triant & DeWoody 2007; Song et al. 2008; Calvignac et al. 2011) and implementing quality control tools such as the use of phylogenetic methods to differentiate nuclear insertions from mitochondrial sequences (Jensen-Seaman et al. 2004; Anthony et al. 2007a).

Finally, the demonstration that several numts are polymorphic in western lowland gorillas and evidence of geographic structuring highlights the potential utility of these insertions as population genetic markers in future studies of gorilla genetic variation. The phylogenetic placement of numt loci in relation to major mitochondrial haplogroups also provides some interesting insights into the divergence history of eastern and western gorillas. Previous studies have shown that the divergence time of western and eastern gorillas and historical gene flow is much more recent that was previously thought (Anthony et al., 2007b; Thalmann et al. 2007; 2011; Scally et al. 2012). Although the class III numt group is sister to the eastern lowland gorilla mitochondrial haplogroup, it is present in both east lowland and western gorillas (Thalmann et al. 2004), suggesting that the transmission of this numt copy to western gorillas occurred relatively recently. This hypothesis is also backed up by the presence of class II numts in eastern lowland and western gorillas. Remarkably however, mtDNA haplogroups are never shared between eastern and western gorillas, providing support for male mediated east-to-west gene flow (Thalmann et al. 2004; 2007) and greater philopatry in female gorillas (Douadi et al. 2007). Combined analyses of craniometric and mitochondrial variation have also provided compelling evidence of ancestral gene flow between western and eastern gorillas and between eastern gorilla subspecies (Ackermann & Bishop 2009). Future work should attempt to relate the distribution of numt polymorphisms to the distribution of major mitochondrial haplogroups in geo-referenced DNA samples from natural populations in order to assess whether patterns of gene flow inferred from presence/absence of numts reflect what is known of gorilla phylogeographic history.

GENERAL CONCLUSIONS

The results from this dissertation have revealed a heterogeneous rate of numt insertion in great apes that varies even between closely related taxa. This observation may be explained by factors such as historical demographic events although differences in the mechanisms that mediate integration and duplication of mitochondrial fragments in the nucleus cannot be ignored. However, this study highlights an underestimation of the rate of insertion relative to previous studies due to a failure to identify numts from highly divergent mitochondrial regions such as MCR and genes under variable or relaxed selection. Comparative analysis of nuclear and mitochondrial paralogous sequences also found evidence of different patterns of nucleotide substitution between mtDNA and their nuclear copies. Interestingly, the potential for numts to behave as “nuclear fossils” was demonstrated as they seem to retain the GC composition of mitochondrial sequences at the time of the translocation to nucleus, thus revealing signs of a gradual historical process of GC content accumulation in the mitochondrial but not the nuclear genome of great apes. A contrast of orthologous genomic regions representing sequences before and after numt insertion provided evidence that rejects a significant role of TEs in the pathway of numt integration and/or duplication while favoring NHEJR as a recurrent mechanism of numt integration. Lastly, the integrated use of bioinformatic searches and experimental isolation of numt sequences in gorillas unveiled several recent mitochondrial insertions in this taxon including several loci with insertional polymorphism that could be used as population genetic markers in future studies. Chapters 2, 3 and 4 describe and discuss in detail the context that led to these findings, which are condensed in the following paragraphs.

I initially compiled a comprehensive database of numts derived from the MCR and present in reference sequenced genomes of humans, chimpanzees and orangutans. A comparison of numt prevalence for each nucleotide position within the MCR and flanking regions revealed a sharp heterogeneity in the number of numts between different MCR subdomains that was partially explained by the distribution of variable sites and unstable repetitive motifs in the MCR. This shows that the apparent deficit of numts from particular MCR subdomains such as HV2 and MCR_F, and to a lesser extent HV1, is largely due to rapid loss of sequence identity between mitochondrial and nuclear paralogs. However, an analysis of numt loci that inserted before and after the origin of Hominoidea, revealed that whereas the deficit of numts derived from HV2 and MCR_F was independent of time of insertion, the deficit of numts from HV1 was only evident in older numts. This indicates that substitution rate in HV2 and

MCR_F has been so high relative to HV1 during the evolution of Hominoidea as to blur the identity of recent insertions. In contrast, substitution rate of HV1 seems to have been lower than the two other subdomains in recent times. Although the effect of differential rates of translocation to nucleus between mitochondrial regions on the observed differences in numt prevalence was not directly assessed, the fact that numts are underestimated due to the loss of sequence divergence is supported by data in this study. Such a hypothesis is also supported by the pattern obtained from the inverse relationship observed between K_a/K_s and the number of numts from protein-coding genes, in that those genes under stronger purifying selection (low K_a/K_s), and therefore more conserved, exhibit a larger number of numts. These lines of evidence from MCR and protein coding mitochondrial genes, added to the fact that this study identified previously unreported numt sequences, also highlights an underestimation in the rate of numt insertion reported in previous studies.

Notable differences in the rate of numt formation were found between great ape taxa. The two most extreme values were found in humans and chimpanzees, which curiously are two of the most recently diverged great ape taxa. The rate in chimpanzees was over 12 fold larger than the estimated value in humans. This result not only points to an excess of numts in chimpanzees but also suggests a great permeability of this genome to accept new integrations, as shown by the great extent of segmental duplications and high rate of structural mutation of this genome (Ventura et al. 2011). Demographic factors have probably played a crucial role in the differences between humans and chimpanzees. It is well established that human populations have been constrained by a historical bottleneck that has eroded genetic diversity as evidenced by reduced levels of genetic variation in several neutral markers (Zhao et al. 2000; Kaessmann et al. 2001; Mathews et al. 2003; Gherman et al. 2007; McEvoy et al. 2011). This is evidenced by data presented here as species with relatively high historical effective population sizes such as chimpanzees and orangutans also surpass humans in the number of numts. Therefore other great apes with larger effective sizes such as gorillas could also have fixed more numts than the observed number in humans (Kaessmann et al. 2001).

Estimates of approximate numt age, sequence comparison of orthologous numts across taxa and an examination of the extent of divergence between contemporary mitochondrial sequences and their nuclear copies provided an unrivaled opportunity to conduct a retrospective assessment of the relative effect of genomic context on the evolution of homologous sequences. It has been hypothesized that mutational bias in the mitochondrial genome has led to substantial increase in the GC content of mitochondrial genomes in great apes (Schmitz et al.

2002; Gibson et al. 2004). Such phenomenon is evidenced by the elevated GC content of contemporary mitochondrial genomes relative to numt sequences. Once a mitochondrial fragment migrates to the nucleus it is thought to escape from the organellar mutational pressure and behave as a “molecular fossil” of the original mitochondrial sequence that retains its properties and nucleotide composition (Brown et al. 1982; Graur & Li 2000; Haag-Liautard et al. 2008). Such a property of numts was demonstrated here after comparisons of numts sequences of varying ages with their corresponding mitochondrial sequences in contemporary genomes. In fact, the difference in GC content between nuclear and mitochondrial sequences increased with numt age, which reveals a non-random variation in the GC content of the mitochondrial genome with a trend towards an excess of GC. These results underscore the potential utility of numts as “molecular fossils”, not only in rooting phylogenies as shown elsewhere (Perna & Kocher 1996; Zischler et al. 1995a,b) but also as witnesses of functional evolutionary changes in the mitochondrial genome.

As expected, T_s/T_v ratios differed significantly between mitochondrial and nuclear sequences as a consequence of differences in patterns and rates of nucleotide substitution. The observed T_s/T_v ratio in mtDNA showed a negative relationship with time of divergence between taxa, a characteristic signature of saturation in the number of transitions at higher levels of genetic divergence that results in underestimation of T_s/T_v ratios (Yang & Yoder 1999). In contrast, T_s/T_v ratio in numts barely changed with time of divergence and remained within the range previously reported for other non-coding regions (Zhang et al. 2007). Although numts derived from coding and non-coding (MCR) mitochondrial regions are both non-functional and randomly insert in all the chromosomes, they differ in patterns of nucleotide substitution, presumably as a consequence of their differences in the intrinsic nucleotide composition. It has been shown for instance that the probability of a given site in the nuclear genome undergoing a transition may be boosted by elevated levels of regional GC content and presence of GC dinucleotides, a phenomenon called “neighbor effect” (Bulmer et al. 1986, Hess et al. 1994). In fact, results in the present study indicate that MCR in great apes has a larger GC content than the mitochondrial average. As a consequence, MCR numts are expected to inherit an elevated GC content that could account for the greater observed T_s/T_v ratio in MCR numts as compared to non-MCR numts.

Past research has shown that the nuclear integration of mitochondrial fragments is most likely opportunistic and mediated by NHEJR in genomic locations with double strand breaks (Blanchard & Schmidt 1996). The contribution that the presented study made is in the use of a large database of numts of known approximate age to simultaneously compare the potential role of NHEJR, *trans*-replication slippage and TE-mediated duplication on numt formation. Identification of nucleotide microhomologies between the numt flank and the inserted mitochondrial fragment in nearly half of the analyzed junction sites was consistent with a role of NHEJR in the mitochondrial integration process. Since presence of microhomologies is not a necessary condition for NHEJR, the role of this mechanism in numt integration may be even more important than previously realized. Comparison of numts made up of non-adjacent mitochondrial fragments also showed substantial presence of microhomologies between tandemly arranged fragments suggesting that NHEJR is also involved in the ligation of these fragments during their integration into the nuclear genome. Lastly, experimental evidence has shown that mitochondrial integration has also occurred through *trans*-replication slippage, but this study failed to find evidence of this mechanism which seems to act under very specific conditions (Chen et al. 2005a). Also, duplication of genomic regions through L1 3' transduction has been previously shown (Goodier *et al.* 2000; Deininger *et al.* 2003). Although this mechanism may theoretically promote numt duplication, experimental evidence presented here indicates that this mode of numt duplication, if possible, would be infrequent and would not represent a predominant way of nuclear colonization of mitochondrial-like sequences.

In the fourth chapter, I focused on the identification of gorilla specific numts and their characterization in a sample of captive gorillas. This is the first study specifically designed to isolate, map and characterize gorilla numts following an integration of experimental and bioinformatic data. The fact that three out of five numts successfully amplified in gorillas and were found to be polymorphic certainly demonstrates that mitochondrial integration is an ongoing process in gorillas. Three numts were also isolated through only one method, highlighting the importance of making use of several complementary methods of numt identification. Four numts were identified through BLAST searches in the short genomic readings used to assemble the recently released draft of the gorilla genome (Scally et al. 2012). Future completion of this genome assembly will probably uncover further insertions and will enable comparable estimations of numt prevalence and insertion rates across all major great ape taxa using the methods described in chapter 2.

Substantial sequence similarity and phylogenetic grouping was found between the gorilla numt classes I and IIb and the numt loci numt2_1 and numt1_1, respectively. This finding provides the first direct validation of two of the three major groups of numts exclusively present in gorillas. Experimental verification of class III numts was not possible in this study despite the use of alternative methods designed to isolate all three numt classes. Although class III numts have a wide geographical distribution as determined from their amplification in both eastern and western lowland gorillas (Anthony et al. 2007a), they have only been found in a limited number of samples suggesting that their frequency might be low, thus reducing the chance of finding them in population samples. Sequencing of numt1_1 and numt2_1 revealed that these two numts comprise the entire HVI region and share extensive sequence identity with authentic HVI haplotypes. Similarity also extends into the annealing regions of traditional HVI primers, to the point that specific amplification of authentic HVI sequences in gorillas is virtually impossible with such primers. Whenever possible, alternative tools designed to circumvent amplification of numts should be used to amplify HVI, including mtDNA isolation and long-range PCR.

In contrast to mtDNA, specific amplification of gorilla numt loci is now possible since specific primers have been designed and tested. Numt presence was only tested in a sample of captive western gorillas whose ancestral geographic origin is not well known. But polymorphic numts described in this research can be amplified in natural populations providing a set of valuable molecular tools to be used in reconstructing historical patterns of gorilla dispersal and hybridization.

REFERENCES

Abeliovich H. 2007. Mitophagy: The life-or-death dichotomy includes yeast. Autophagy. 3: 275-277.

Ackermann RR and Bishop JM. 2009. Morphological and molecular evidence reveals recent hybridization between gorilla taxa. Evolution. 64: 271-290.

Adams KL and Palmer JD. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol. Phylog. Evol. 29: 380-395.

Adams KL, Daley DO, Qiu, YL, Whelan J, Palmer JD. 2000. Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature. 408: 354-357.

Ahmed ZM, Smith TN, Riazuddin S, Makishima T, Ghosh M, Bokhari S, Menon P, Deshmukh D, Griffith A, Riazuddin S, Friedman T, Wilcox E. 2002. Nonsyndromic recessive deafness DFNB18 and Usher syndrome type IC are allelic mutations of USHIC. Hum. Genet. 110: 527–531.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215: 403-410.

Anthony NM, Clifford SL, Bawe-Johnson M, Abernethy KA, Bruford MW, Wickings EJ. 2007a. Distinguishing gorilla mitochondrial sequences from nuclear integrations and PCR recombinants: Guidelines for their diagnosis in complex sequence databases. Mol. Phylog. Evol. 43: 553-566.

Anthony NM, Johnson-Bawe M, Jeffery K, Clifford SL, Abernethy KA, Tutin CE, Lahm SA, White LJ, Utley JF, Wickings EJ, Bruford MW. 2007b. The role of Pleistocene refugia and rivers in shaping gorilla genetic diversity in central Africa. Proc. Natl. Acad. Sci USA. 104: 20432-20436.

Antunes A and Ramos MJ. 2005. Discovery of a large number of previously unrecognized mitochondrial pseudogenes in fish genomes. Genomics. 86: 708-717.

Antunes A, Pontius J, Ramos MJ, O’Brien SJ, Johnson WE. 2007.Mitochondrial introgressions into the nuclear genome of the domestic cat. J. Hered. 98: 414–420.

Arctander P. 1995. Comparison of a mitochondrial gene and a corresponding nuclear pseudogene. Proc. R. Soc. Lond. B. 262: 13-19.

Arndt PF, Petrov DA and Hwa T. 2003. Distinct Changes of Genomic Biases in Nucleotide substitution at the Time of Mammalian Radiation. Mol. Biol. Evol. 20: 1887-1896.

Arora N, Nater A, van Schaik CP, Willems EP, van Noordwijk MA, Goossens B, Morf N, Bastian M, Knott C, Morrogh-Bernard H, Kuze N, Kanamori T, Pamungkas J, Perwitasari-Farajallah D, Verschoor E, Warren K, Krützen M. 2010. Effects of Pleistocene glaciations and rivers on the population structure of Bornean orangutans (*Pongo pygmaeus*). Proc. Natl. Acad. Sci. USA. 21376-21381.

Ashley N, Harris D, Poulton J. 2005. Detection of mitochondrial DNA depletion in living human cells using PicoGreen staining. Experimental Cell Res. 303: 432-446.

Batzer MA and Deininger PL. 2002. Alu repeats and human genetic diversity. Nature Rev. Genet. 3: 370-380.

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, Novick GE, Ioannou PA, Scheer WD and Herrera RJ.1994. African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci USA. 91: 12288-12292.

Behura SK. 2007. Analysis of nuclear copies of mitochondrial sequences in honeybee (*Apis mellifera*) genome. Mol. Biol. Evol. 24: 1492–1505.

Bensasson D, Feldman MW, Petrov DA. 2003. Rates of DNA Duplication and Mitochondrial DNA Insertion in the Human Genome. J. Mol. Evol. 57: 343-354.

Bensasson D, Zhang D, Hartl DL and Hewitt GM. 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. Trends Ecol.Evol. 16: 314-321.

Bensasson D, Zhang DX, Hewitt GM. 2000. Frequent assimilation of mitochondrial DNA by grasshopper nuclear genomes. Mol. Biol. Evol. 17: 406-415.

Berg OG and Kurland CG. 2000. Why mitochondrial genes are most often found in nuclei. Mol. Biol. Evol. 17: 951-961.

Berger A, Bruschek M, Grethen C, Sperl W, Kofler B. 2001. Poor storage and handling of tissue mimics mitochondrial DNA depletion. Diagn. Mol. Pathol. 10: 55-59.

Bertheau C, Schuler H, Krumböck S, Arthofer W, Stauffer C. 2011. Hit or miss in phylogeographic analyses: the case of the cryptic NUMTs. Mol. Ecol. Res. 11: 1056-1059.

Black IV WC and Bernhardt SA. 2009. Abundant nuclear copies of mitochondrial origin (NUMTs) in the *Aedes aegypti* genome. Insect Mol. Biol. 18: 705-713.

Blanchard JL and Lynch M. 2000. Organellar genes: Why do they end up in the nucleus? Trends Genet. 16: 315-320.

Blanchard JL and Schmidt GW. 1996. Mitochondrial DNA Migration Events in Yeast and Humans: Integration by a Common End-Joining Mechanism and Alternative Perspectives on Nucleotide Substitution Patterns. Mol. Biol. Evol. 13: 537-548.

Bodenteich A, Mitchell LG, Polymeropoulos MH and Merril CR. 1992. Dinucleotide repeat in the human mitochondrial D-loop. Hum. Mol. Genet. 1:140.

Boore JL. 1999. Animal mitochondrial genomes. Nucleic Acid Research. 27: 1767-1780.

Borensztajn K, Chafa O, Alhenc-Gelas M, Salha S, Reghis A, Fischer AM, Tapon-Bretaudière J. 2002. Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. British J. Haematol. 117: 168-171.

Broughton RE and Reneau PC. 2006. Spatial Covariation of Mutation and Nonsynonymous Substitution Rates in Vertebrate Mitochondrial Genomes. Mol. Biol.Evol. 23:1516–1524.

Brown WM, Prager EM, Wang A and Wilson AC. 1982. Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J. Mol. Evol. 18: 225-239.

Calloway CD, Reynolds RL, Herrin GL, Anderson WW. 2000. The Frequency of Heteroplasmy in the HVII Region of mtDNA Differs across Tissue Types and Increases with Age. Am. J. Hum. Genet. 66: 1384-1397.

Calvignac S, Konecny L, Malard F, Douady CJ. 2011. Preventing the pollution of mitochondrial datasets with nuclear mitochondrial paralogs (numts). Mitochondrion. 11: 246-254.

Campbell CL and Thorsness PE. 1998. Escape of mitochondrial DNA to the nucleus in yme1 yeast is mediated by vacuolar-dependent turnover of abnormal mitochondrial compartments. J. Cell Science. 111: 2455-2464.

Castella V, Dimo-Simonin N, Brandt-Casadevall C, Robinson N, Saugy M, Taroni F, Mangin P. 2006. Forensic identification of urine samples: a comparison between nuclear and mitochondrial DNA markers. Int. J. Leg. Med. 120: 67-72.

Chen JM, Chuzhanova N, Stenson PD, Férec C and Cooper DN. 2005. Meta-analysis of gross insertions causing human genetic disease: Novel mutational mechanisms and the role of replication slippage. Hum. Mut. 25: 207-221.

Chinnery PF, Thourburn DR. Samuels DC, White SL, Dahl HHM, Turnbull DM, Lightowlers RN, Howell N. 2000. The inheritance of mitochondrial DNA heteroplasmy: Random drift, selection or both? Trends in Genet. 16: 500-505.

Chung W and Steiper M. 2008. Mitochondrial COII Introgression into the Nuclear Genome of Gorilla gorilla. Int. J. Primat. 29: 1341-1353.

Clifford SL, Anthony NM, Bawe-Johnson M, Abernethy KA, Tutin CEG, White LJT, Bermejo M, Goldsmith ML, Mcfarland K, Jeffery KJ, Bruford MW, Wickings EJ. 2004. Mitochondrial DNA phylogeography of western lowland gorillas (*Gorilla gorilla gorilla*). Mol. Ecol. 13: 1551-1565.

Collura RV and Stewart CB. 1995. Insertions and duplications of mtDNA in the nuclear genomes of Old World monkeys and hominoids. Nature. 378: 485–489.

Corral M, Baffet G, Kitzis A, Paris B, Tichonicky L, Kruh J, Guguen-Guillouzo C, Defer N. 1989. DNA sequences homologous to mitochondrial genes in nuclei from normal rat tissues and from rat hepatoma cells. Biochem Biophys Research Comm. 162: 258-264.

de Grey ADNJ. 2005. Forces maintaining organellar genomes: is any as strong as genetic code disparity or hydrophobicity? BioEssays 27: 436-446.

Deininger PL and Batzer MA. 2002. Mammalian retroelements. Genome Res. 12: 1455-14-65.

Deininger PL, Morany JV, Batzer MA and Kazazian Jr HH. 2003. Mobile elements and mammalian genome evolution. Curr. Opinion Genet. Develop. 13: 651-658.

DeWoody JA, Chesser RK, Baker RJ. 1999. A Translocated Mitochondrial Cytochrome b Pseudogene in Voles (Rodentia: *Microtus*). J. Mol. Evol. 48: 380-382.

Douadi MI, Gatti S, Levrero F, Duhamel G, Bermejo M, Vallet D, Menard N, Petit EJ. 2007. Sex-biased dispersal in western lowland gorillas (*Gorilla gorilla gorilla*). Mol. Ecol. 16: 2247-2259.

Du Buy HG and Riley FL. 1967. Hybridization between the nuclear and kinetoplast DNA's of Leishmania enriettii and between nuclear and mitochondrial DNA's of mouse liver. Proc. Natl. Acad. Sci. USA. 57: 790-797.

Erlandsson R, Wilson JF and Pääbo S. 2000. Sex chromosomal transposable element accumulation and male-driven substitutional evolution in humans. Mol. Biol. Evol. 17: 804-812.

Farrelly F and Butow RA. 1983. Rearranged mitochondrial genes in the yeast nuclear genome. Nature. 301: 296-301.

Fernandez-Silva P, Enriquez JA, Montoya J. 2003. Replication and transcription of mammalian mitochondrial DNA. Exp. Physiol. 88: 41-56.

Foote S, Vollrath D, Hilton A and Page DC. 1992. The human Y chromosome: Overlapping DNA clones spanning the euchromatic region. Science 258: 60-66.

Foran DR. 2006. Relative degradation of nuclear and mitochondrial DNA: An experimental approach. J. Forensic. Sci. 51: 766-770.

Frey JE and Frey B. 2004. Origin of intra-individual variation in PCR-amplified mitochondrial cytochrome oxidase I of *Thrips tabaci* (Thysanoptera: Thripidae): mitochondrial heteroplasmy or nuclear integration? Hereditas. 140: 92-98.

Gagneux P, Wills C, Gerloff U, Tautz D, Morin PA, Boesch C, Fruth B, Hohmann G, Ryder OA, Woodruff DS. Mitochondrial sequences show diverse evolutionary histories of African hominoids. Proc. Natl. Acad. Sci.USA. 96: 5077-5082.

Garner KJ and Ryder OA. 1996. Mitochondrial DNA diversity in gorillas. Mol. Phylog. Evol. 6: 39-48.

Gherman A, Chen PE, Teslovich TM, Stankiewicz P, Withers M, Kashuk CS, Chakravarti A, Lupski JR, Cutler DJ, Katsanis N. 2007. Population bottlenecks as a potential major shaping force of human genome architecture. PLoS 3 (7): e119.

Giampieri C, Centurelli M, Bonafè M, Olivieri F, Cardelli M, Marchegiani F, Cavallone LGiovagnetti S, et al. 2004. A novel mitochondrial DNA-like sequence insertion polymorphism in Intron I of the FOXO1A gene. Gene. 327: 215–219.

Gibson et al. 2004. A Comprehensive Analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol. Biol. Evol. 22: 251–264.

Goldin, E, Stahl S, Cooney AM, Kaneski CR, Gupta S, Brady RO, Ellis JR, Schiffmann R. 2004. Transfer of a mitochondrial DNA fragment to MCOLN1 causes an inherited case of mucolipidosis IV. Hum. Mut. 24: 460-465.

Goldman SJ, Taylor R, Zhang Y, Shengkan J. 2010. Autophagy and the degradation of mitochondria. Mitochondrion 10: 309-315.

Goodier JL, Ostertag EM and Kazaziar Jr. HH. 2000. Transduction of 3’-flanking sequences is common in L1 retrotransposition. Hum. Mol. Genet. 9: 653-657.

Goodman M, Porter CA, Czelusniak J, Page SL, Schneider H, Shoshani J, et al. 1998. Toward a phylogenetic classification of primates based on DNA evidence complemented by fossil evidence. Mol. Phylogenet. Evol. 9:585-598.

Graur D & Li WH. 2000. Fundamentals of Molecular Evolution. 2nd Edition. Sinauer Associates, Inc. Sunderland, MA, USA.

Gray MW and Doolittle WF. 1982. Has the endosymbiont hypothesis been proven? Microbiol. Rev. 46: 1-42.

Gray MW. 1999. Evolution of organellar genomes. Curr. Opin. Genet. Dev. 9: 678-687.

Greenwood AD and Päävo S. 1999. Nuclear insertion sequences of mitochondrial DNA predominate in hair but not in blood of elephants. Mol. Ecol. 8: 133-137.

Grosso AR, Basto-Silveira C, Coelho MM, Dias D. 2006. *Columba palumbus* Cyt b-like Numt sequence: comparison with functional homologue and the use of universal primers. Folia Zoologica. 55: 131-144.

Grubb P, Butynski TM, Oates JF, Bearder SK, Disotell TR, Groves CP, Struhsaker TT. 2003. Assessment of the diversity of African primates. Int. J. Primat. 24: 1301-1357.

Gyllensten U, Wharton D, Josefsson A, Wilson AC. 1991. Paternal inheritance of mitochondrial DNA in mice. Nature. 352: 255-257.

Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B and Keightley PD. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. PLoS Biol 6: e204. doi:10.1371/journal.pbio.0060204

Hay JM, Sarre SD, Daugherty CH. 2004. Nuclear mitochondrial pseudogenes as molecular outgroups for phylogenetically isolated taxa: a case study in *Sphenodon*. Heredity. 93: 468-475.

Hazkani-Covo E and Graur D. 2007. A comparative analysis of numt evolution in human and chimpanzee. Mol. Biol. Evol. 24: 13-18.

Hazkani-Covo E and Covo S. 2008. Numt-Mediated double-strand break repair mitigates deletions during primate genome evolution. Plos Genetics. 40: e1000237. doi:10.1371/journal.pgen.1000237.

Hazkani-Covo E, Sorek R, Graur D. 2003. Evolutionary dynamics of large Numts in the human genome: rarity of independent insertions and abundance of post-insertion duplications. J. Mol. Evol. 56: 169-174.

Hazkani-Covo E, Zeller RM, Martin W. 2010. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. PLoS Genet 6: e1000834. doi:10.1371/journal.pgen.1000834.

Hazkani-Covo E. 2009. Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. Mol. Biol. Evol. 26: 2175-2179.

Hedges SB and Schweitzer MH. 1995. Detecting dinosaur DNA. Science 268: 1191-1192

Henze K and Martin W. 2001. How do mitochondrial genes get into the nucleus? Trends. Genet. 17: 383-387.

Herke SW, Xing J, Ray DA, Zimmerman JW, Cordaux R, Batzer MA. 2007. A SINE-based dichotomous key for primate identification. Gene. 390: 39-51.

Hirano M, Shtilbans A, Mayeux R, Davidson MM, DiMauro S, Knowles JA, Schon EA. 1997. Apparent mtDNA heteroplasmy in Alzheimer’s disease patients and in normals due to PCR amplification of nucleus-embedded mtDNA pseudogenes. Proc. Natl. Acad. Sci. USA. 94: 14894-14899.

Hudson ME. 2008. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol. Ecol. Res. 8: 3-17.

Ibarguchi G, Friesen VL, Lougheed SC. 2006. Defeating numts: Semi-pure mitochondrial DNA from eggs and simple purification methods for field-collected wildlife tissues. Genome. 49: 1438-1450.

Ingman M and Gyllensten U. 2006. MtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acid. Res. 34: D749–D751.

Jackson SP. 2002. Sensing and repairing DNA double-strand breaks. Carcinogenesis. 23: 687-696.

Jensen-Seaman MI and Kidd KK. 2001. Mitochondrial DNA variation and biogeography of eastern gorillas. Mol. Ecol. 10: 2241-2247.

Jensen-Seaman MI, Wildschutte JH, Soto-Calderón ID, Anthony NM 2009. A comparative approach reveals differences in patterns of numt insertion during hominoid evolution. J. Mol. Evol. 68: 688-699.

Jensen-Seaman MJ, Sarmiento EE, Deinard A.S, Kidd KK. 2004. Nuclear integrations of mitochondrial DNA in gorillas. Am. J. Primat. 63: 139-147.

Jenuth JP, Peterson AC, Shoubridge EA. 1997. Tissue-specific selection for different mtDNA genotypes in heteroplasmic mice. Nature Genetics. 16: 93-95.

Jones DH and Winistorfer SC. 1992. Sequence specific generation of a DNA panhardle permits PCR amplification of unknown flanking DNA. Nucleic Acids Res. 20: 596-600.

Judo MSB, Wedel AB, Wilson C. 1998. Stimulation and suppression of PCR-mediated recombination. Nucleic Acids Res. 26: 1819-1825.

Kaessmann H, Wiebe V, Weiss G and Pääbo S. 2001. Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nature 27: 155-156.

Kamimura N, Ishii S, Ma LD, Shay JW. 1989. Three separate mitochondrial DNA sequences are contiguous in human genomic DNA. J. Mol Biol. 210: 703–707.

Kanki T and Klionsky DJ. 2010. The molecular mechanism of mitochondria autophagy in yeast. Mol. Microbiol. 75: 795-800.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. Genome Res. 12: 996-1006.

Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica. 115: 49-63.

Kocher TD, Thomas WK, Meyer A, Edwards SV, Pääbo S, Villablanca FX, Wilson AC. 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA. 86: 6196-6200.

Kondo R, Satta Y, Matsuura ET, Ishiwa H, Takahata N, Chigusa SI. 1990. Incomplete maternal Transmission of Mitochondrial DNA in *Drosophila*. Genetics. 126: 657-663.

Kvist L, Martens J, Nazarenko AA, Orell M. 2003. Paternal Leakage of Mitochondrial DNA in the Great Tit (*Parus major*). Mol. Biol. Evol. 20: 243-247.

Lacoste V, Mauclère P, Dubreuil G, Lewis J, Georges-Courbot MC, Gessain A. 2001. A Novel γ2-herpesvirus of the rhadinovirus 2 lineage in chimpanzees. Genome Res. 11: 1511-1519.

Lacoste V, Mauclère P, Dubreuil G, Lewis J, Georges-Courbot MC, Gessain A. 2001. A Novel γ2-herpesvirus of the rhadinovirus 2 lineage in chimpanzees. Genome Res. 11: 1511-1519.

Lang M, Sazzini M, Calabrese F, Simone D, Boattini A, Romeo G, Luiselli D, Attimonelli M, Gasparre G. 2011. Polymorphic NumtS trace human population relationships. Hum. Genet. 1511-1519.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ and Higgins DG. 2007. ClustalW and ClustalX version 2. Bioinformatics 23: 2947-2948.

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C and Attimonelli M. 2008. The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify human NumtS. BMC Genomics. 9: 267.

Lascaro D, Castellana S, Gasparre G, Romeo G, Saccone C and Attimonelli M. 2008. The RHNumtS compilation: Features and bioinformatics approaches to locate and quantify human NumtS. BMC Genomics 9: 267.

Leister D. 2005. Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends Genet. 21: 655-663.

Liang BC. 1996. Evidence for association of mitochondrial DNA sequence amplification and nuclear localization in human low-grade gliomas. Mutation Res. 354: 27-33.

Librado P and Rozas J. 2009. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. Bioinformatics. 25: 1451-1452.

Lopez JV, Cevario S, O'Brien SJ. 1996. Complete Nucleotide Sequences of the Domestic Cat (*Felis catus*) Mitochondrial Genome and a Transposed mtDNA Tandem Repeat (Numt) in the Nuclear Genome. Genomics. 33: 229-246.

Lopez JV, Culver M, Stephens JC, Johnson WE and O'brien SJ. 1997. Rates of nuclear and cytoplasmic mitochondrial DNA sequence divergence in mammals. Mo. Biol. Evol. 14: 277-286.

Lopez JV, Yuhki N, Masuda R, Modi W and O'Brien SJ. 1994. Numt, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. J. Mol. Evol. 39: 174-190.

Malek O, Brennicke A, Knoop V. 1997. Evolution of trans-splicing plant mitochondrial introns in pre-Permian times. Proc. Natl. Acad. Sci. USA. 94: 553-558.

Mardis ER. 2008. Next-Generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9: 387-402.

Mason VC, Li G, Helgen KM, Murphy WJ. 2011. Efficient cross-species capture hybridization and next-generation sequencing of mitochondrial genomes from noninvasively sampled museum specimens. Genome Res. 21: 1695-1704.

Mathews LM, Chi SY, Greenberg N, Ovchinnikov I and Swergold GD. 2003. Large Differences between LINE-1 amplification rates in the human and chimpanzee lineages. Am. J. Hum. Genet. 72: 739-748.

McEvoy BP, Powell JE, Goddard ME, Visscher PM. 2011. Human population dispersal ''Out of Africa'' estimated from linkage disequilibrium and allele frequencies of SNPs. Genome Res. 21: 821-829.

McLeod BA and White BN. 2010. Tracking mtDNA heteroplasmy through multiple generations in the North Atlantic right whale (*Eubalaena glacialis*). J. Hered. 101: 235-239.

Meiklejohn CD, Montooth KL, Rand DM. 2007. Positive and negative selection on the mitochondrial genome. Trends Genet. 23: 259-263.

Mijaljica D, Prescott M, Devenish RJ. 2007. Different fates of mitochondria: Alternative ways for degradation? Autophagy. 3: 4-9.

Miraldo A, Hewitt GM, Dear PH, Paulo OS, Emerson BC. 2012. Numts help to reconstruct the demographic history of the ocellated lizard (*Lacerta lepida*) in a secondary contact zone. Mol. Ecol. 21: 1005-1018.

Mishmar D, Ruiz-Pesini E, Brandon M and Wallace DC. 2004. Mitochondrial DNA-like sequences in the nucleus (NUMTs): Insights into our African origins and the mechanism of foreign DNA integration. Hum. Mut. 23: 125-133.

MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2008.

Moran JV, DeBerardinis RJ and Kazazian HH Jr. 1999. Exon shuffling by L1 retrotransposition. Science. 283: 1530-1534.

Moulton MJ, Song H, Whiting MF. 2010. Assessing the effects of primer specificity on eliminating numt coamplification in DNA barcoding: A case study from Orthoptera (Arthropoda: Insecta). Mol. Ecol. Res. 10: 615-627.

Mourier T, Hansen AJ, Willerslev E and Arctander P. 2001. The human genome project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol. Biol. Evol. 18: 1833-1837.

Mundy NI, Pissinatti A and Woodruff DS. 2000. Multiple nuclear insertions of mitochondrial cytochrome b sequences in Callitrichine primates. Mol. Biol. Evol. 17: 1075-1080.

Naidu A, Fitak RR, Munguia-Vega A, Culver M. 2012. Novel primers for complete mitochondrial cytochrome b gene sequencing in mammals. Mol. Ecol. Res. 12: 191-196.

Nergadze SG, Lupotto M, Pellanda P, Santagostino M, Vitelli V, Giulotto E. 2010. Mitochondrial DNA insertions in the nuclear horse genome. Animal Genet. 41: 176-185.

Nugent JM and JD Palmer. 1991. RNA-mediated transfer of the gene coxII from the mitochondrion to the nucleus during flowering plant evolution. Cell. 66: 473-481.

Ochman H, Gerber AS, Hartl DL. 1988. Genetic applications of an inverse polymerase chain reaction. Genetics. 120: 621-623.

Pääbo S, Irwin DM, Wilson AC. 1990. DNA Damage promotes jumping between templates during enzymatic amplification*. J. Biol. Chem. 265: 4718-4721.

Pamilo P, Viljakainen L, Vihavainen A. 2007. Exceptionally high density of NUMTs in the honeybee genome. Mol. Biol. Evol. 24: 1340-1346.

Park S, Hanekamp T, Thorsness MK, Thorsness PE. 2006. Yme2p is a mediator of nucleoid structure and number in mitochondria of the yeast *Saccharomyces cerevisiae*. Curr. Genet. 50: 173–182.

Perna NT and Kocher TD. 1996. Mitochondrial DNA: molecular fossils in the nucleus. Curr. Biol. 6: 128-129.

Perna NT, Batzer MA, Deininger PL, Stoneking M.1992. Alu insertion polymorphism: A new type of marker for human population studies. Human Biol. 64: 641-648.

Pesole G, Gissi C, De Chirico A and Saccone C. 1999. Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. J. Mol. Evol. 48: 427-434.

Pickeral OK, Makalowski W, Boguski MS, Boeke JD. 2000. Frequent human genomic DNA transduction driven by LINE-1 retrotransposition. Genome Res. 10: 411–415.

Podnar M, Haring E, Pinsker W, Mayer W. 2007. Unusual Origin of a Nuclear Pseudogene in the Italian Wall Lizard: Intergenomic and Interspecific Transfer of a Large Section of the Mitochondrial Genome in the Genus *Podarcis* (Lacertidae). J. Mol. Evol. 64: 308-320.

Posada D. 2008. jModelTest: Phylogenetic Model Averaging. Mol. Biol. Evol. 25: 1253-1256.

Priault M, Salin B, Schaeffer J, Vallette FM, di Rago JP, Martinou JC. 2005. Impairing the bioenergetic status and the biogenesis of mitochondria triggers mitophagy in yeast. Cell Death Differ. 12: 1613–1621.

Purvis A and Bromham L. 1997. Estimating the Transition/Transversion ratio from independent pairwise comparisons with an assumed phylogeny. J. Mol. Evol. 44: 112-119.

Rawi SA, Louvet-Vallée S, Djeddi A, Sachse M, Culetto E, Hajjar C, Boyd L, Legouis R, Galy V. 2011. Postfertilization autophagy of sperm organelles prevents paternal mitochondrial DNA transmission. Science. 334: 1144-1147.

Ray DA, Xing J, Hedges DJ, Hall MA, Laborde ME, Anders BA, White BR, Stoilova N, Fowlkes JD, Landry KE, Chemnick LG, Ryder OA, Batzer MA. 2005. Alu insertion loci and platyrrhine primate phylogeny. Mol. Phylog. Evol. 35: 117-126.

Ray DA, Xing J, Salem AH, Batzer MA. 2006. SINEs of a Nearly Perfect Character. Syst. Biol. 55: 928-935.

Ren M, Chen Q, Li L, Zhang R, Guo S. 2005. Successive chromosome walking by compatible ends ligation inverse PCR. Mol. Biotech. 30: 95-101.

Ricchetti M, Fairhead C and Dujon B. 1999. Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature. 402: 96-100.

Ricchetti M, Tekaia F and Dujon B. 2004. Continued Colonization of the Human Genome by Mitochondrial DNA. PLOS Biology. 2: 1313-1324. doi: 10.1371/journal.pbio.0020273

Richly E and Leister D. 2004. NUMTs in sequenced eukaryotic genomes. Mol. Biol. Evol. 21: 1081-1084.

Roon DA, Waits LP, Kendall KC. 2003. A quantitative evaluation of two methods for preserving hair samples. Mol. Ecol. Notes. 3: 163.

Roth DB, Porter TN and Wilson JH. 1985. Mechanisms of nonhomolgous recombination in mammalian cells. Mol. Cell. Biol. 5: 2599-2607.

Saccone C, Pesole G and Sbisà. 1991. The main regulatory region of mammalian mitochondrial DNA: Structure-function model and evolutionary pattern. J. Mol. Evol. 33: 83-91.

Saccone S, Federico C, Bernardi G. 2002. Localization of the gene-richest and the gene-poorest isochores in the interphase nuclei of mammals and birds. Gene. 300: 169-178.

Sacerdot C, Casaregola S, Lafontaine I, Tekaia F, Dujon B, Ozier-Kalogeropoulus O. 2008. Promiscuous DNA in the nuclear genomes of hemiascomycetous yeasts. FEMS Yeast Res. 8: 846-857.

Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, Mullis KB, Erlich HA. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 239: 487-491.

Sato M and Sato K. 2011. Degradation of paternal mitochondria by fertilization-triggered autophagy in *C. elegans* embryos. Science. 334: 1141-1144.

Sbisà E, Tanzariello F, Reyes A, Pesole G and Saccone C. 1997. Mammalian mitochondrial D-loop region structural analysis: Identification of new conserved sequences and their functional and evolutionary implications. Gene. 205: 125-140.

Scally A, Dutheil JY, et al. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature. 483: 169-175.

Schmitz J, Ohme M and Zischler H 2002. The complete mitochondrial sequence of Tarsius bancanus: Evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Mol. Biol. Evol 19: 544-553.

Schmitz J, Piskurek O, Zischler H. 2005. Forty million years of independent evolution: A mitochondrial gene and its corresponding nuclear pseudogene in primates. J. Mol. Evol. 61: 1-11.

Schwartz and Vissing 2002. Paternal inheritance of mitochondrial DNA. N. Engl. J. Med. 347: 576-580.

Selosse MA, Albert B, Godelle B. 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol. Evol. 16: 135-141.

Shafer KS, Hanekamp T, White KH, Thorsness PE. 1999. Mechanisms of mitochondrial DNA escape to the nucleus in the yeast *Saccharomyces cerevisiae*. Curr. Genet. 36: 183–194.

Shay JW and Werbin H. 1992. New evidence for the insertion of mitochondrial DNA into the human genome: Significance for cancer and aging. Mutation Res. 275: 227-235.

Smit AFA, Hubley R and Green P. RepeatMasker Open-3.0. http://www.repeatmasker.org. 1996-2007.

Smith MF, Thomas WK, Patton JL. 1992. Mitochondrial DNA-like sequence in the nuclear genome of an Akodontine rodent. Mol. Biol. Evol. 9: 204-215.

Song H, Buhay JE, Whiting MF, Crandall KA. 2008. Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. Proc. Natl. Acad. Sci. USA. 5: 13486–13491.

Soto-Calderón ID, Ntie S, Mickala P, Maisels F, Wickings EJ, Anthony NM. 2009. Effects of storage type and time on DNA amplification success in tropical ungulate faeces. Molec.Ecol. Resour. 9: 471-479.

Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J. 2001. Complex mtDNA constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2: Implication of potential sequencing errors caused by large-unit repeats. Proc. Natl. Acad. Sci. USA. 98: 5099-5103.

Sunnucks P and Hales DF. 1996. Numerous transposed sequences of mitochondrial cytochrome oxidase I-II in aphids of the genus *Sitobion* (Hemiptera: Aphididae). Mol. Biol. Evol. 13: 510-524.

Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G. 1999. Development: Ubiquitin tag for sperm mitochondria. Nature. 402: 371-372.

Swofford DL. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4.0b10 (Alvitec). Sunderland, Massachusetts: Sinauer Associates.

Tamura K, Dudley J, Nei M and Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24: 1596-1599.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2011. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28: 2731-2739.

Tan G, Gao Y, Shi M, Zhang X, He S, Chen Z, An C. 2005. SiteFinding-PCR: a simple and efficient PCR method for chromosome walking. Nucleic Acids Res. 33: e122.

Thalmann O, Fischer A, Lankester F, Pääbo S and Vigilant L. 2007. The complex evolutionary history of gorillas: Insights from genomic data. Mol. Biol. Evol. 24: 146-158.

Thalmann O, Hebler J, Poinar HN, Pääbo S, Vigilant L. 2004. Unreliable mtDNA data due to nuclear insertions: a cautionary tale from analysis of humans and other great apes. Mol. Ecol. 13: 321-335.

Thalmann O, Serre D, Hofreiter M, Lukas D, Eriksson J, Vigilant L. 2005. Nuclear insertions help and hinder inference of the evolutionary history of gorilla mtDNA. Mol. Ecol. 14: 179-188.

Thalmann O, Wegmann D, Spitzner M, Arandjelovic M, Guschanski K, Leuenberger C, Bergl RA, Vigilant L. 2011. Historical sampling reveals dramatic demographic changes in western gorilla populations. BMC Evol. Biol. 11: 85.

Thomas R, Zischler H, Päävo S, Stoneking M. 1996. Novel mitochondrial DNA insertion polymorphism and its usefulness for human population studies. Hum. Biol. 68: 847-854.

Thorness PE and Fox TD. 1993. Nuclear mutations in *Saccharomyces cerevisiae* that affect the escape of DNA from mitochondria to the nucleus. Genetics. 134: 21-28.

Thorsness PE Weber ER. 1996. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. Int. Rev. Cytol. 165: 207-234.

Thurston MI and Field D. 2005. Msatfinder: detection and characterisation of microsatellites. Available from: http://www.bioinf.ceh.ac.uk/msatfinder/.

Tilford CA, Kuroda-Kawagushi T, Skaletsky H, Rozen S, Brown LG, Rosenberg M, McPherson JD, Wylie K, Sekhon M, Kucaba TA, Waterson RH and Page DC. 2001. A physical map of the human Y chromosome. Nature. 409: 943-945.

Timmis JN, Ayliffe MA, Huang CY and Martin W. 2004. Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. Nature Rev. Genet. 5: 123-134.

Tourmen Y, Baris O, Dessen P, Jacques C, Malthièry Y, Reynier P. 2002. Structure and chromosomal distribution of human mitochondrial pseudogenes. Genomics. 80: 71-77.

Triant DT and DeWoody JA. 2007. The occurrence, detection, and avoidance of mitochondrial DNA translocations in mammalian systematics and phylogeography. J. Mammal. 88:908-929.

Turner C, Killoran, Thomas NST, Rosenberg M, Chuzhanova NA, Johnston J, Kemel Y, Cooper DN and Biesecker LG. 2003. Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. Hum. Genet. 112: 303-309.

Twig G, Elorza A, Molina AJA, Mohamed H, Wikstrom JD, Walzer G, Stiles L, Haigh SE, Katz S, Las G, Alroy J, Wu M, Py BF, Yuan J, Deeney JT, Corkey BE, Shirihai OS. 2008a. Fission and selective fusion govern mitochondrial segregation and elimination by autophagy. EMBO J. 27: 433-446.

Twig G, Hyde B, Shirihai OS. 2008b. Mitochondrial fusion, fission and autophagy as a quality control axis: The bioenergetic view. Biochim. Biophys. Acta. 1777: 1092-1097.

Vallinoto M, Sena L, Sampaio I, Schneider H, Schneider MP. 2000. Mitochondrial DNA-like sequence in the nuclear genome of *Saguinus* (Callitrichinae, Primates): Transfer estimation. Genet. Mol. Biol. 23: 35-42.

van den Boogaart P, Samallo J. Agsteribbe E. 1982. Similar genes for a mitochondrial ATPase subunit in the nuclear and mitochondrial genomes of *Neurospora crassa*. Nature. 298: 187-189.

van der Kuyl AC, Kuiken CL, Dekker JT, Perizonius WR, Goudsmit J. 1995. Nuclear counterparts of the cytoplasmic mitochondrial 12S rRNA gene: a problem of ancient DNA and molecular phylogenies. J. Mol. Evol. 40: 652-657.

Vartanian JP and Wain-Hobson S. 2002. Analysis of a library of macaque nuclear mitochondrial sequences confirms macaque origin of divergent sequences from old oral polio vaccine samples. Proc. Natl. Acad. Sci. USA. 99: 7566-7569.

Ventura M, Catacchio CR, Alkan C, Marques-Bonet T, Sajjadian S, Graves TA, Hormozdiari F, Navarro A, Malig M, Baker C, Lee C, Turner EH, Chen L, Kidd JM, Archidiacono N, Shendure J, Wilson RK, Eichler EE. 2011. Genome Res. 21: 1640-1649.

Walker JA, Hughes DA, Anders BA, Shewale J, Sinha SK, Batzer MA. 2003. Quantitative intra-short interspersed element PCR for species-specific DNA identification. Analytical Biochem. 316: 259-269.

Wallace DC, Studgard C, Murdock D, Schurr T and Brown M.D. 1997. Ancient mtDNA sequences in the human nuclear genome: A potential source of errors in identifying pathogenic mutations. Proc. Natl. Acad. Sci. USA. 94: 14900-14905.

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AME, Carroll ML, Nguyen SV, Walker JA, Prasad BVR, Reddy PG, Das PK, Batzer MA, Jorde LB. 2003. Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. Genome Res. 13: 1607-1618.

Wharton D. 2007. North American studbook for the western lowland gorilla (*Gorilla gorilla gorilla*). Chicago Zoological Society. Brookfield, IL, USA.

Willett-Brozick JE, Savul SA, Richey LE and Baysal BE. 2001. Germ line insertion of mtDNA at the breakpoint junction of a reciprocal constitutional translocation. Hum. Genet. 109: 216-223.

Williams ST and Knowlton N. 2001. Mitochondrial pseudogenes are pervasive and often insidious in the snapping shrimp genus *Alpheus*. Mol. Biol. Evol. 18: 1484-1493.

Woischnik M and Moraes CT. 2002. Pattern of Organization of Human Mitochondrial Pseudogenes in the nuclear genome. Genome Res. 12: 885-893.

Xing J, Wang H, Belancio VP, Cordaux R, Deininger PL, Batzer MA. 2006. Emergence of primate genes by retrotransposons-mediated sequence transduction. Proc. Natl. Acad. Sci. USA. 103: 17608-17613.

Yang Z and Yoder AD. 1999. Estimation of the Transition/Transversion rate bias and species sampling. J. Mol. Evol. 48: 274-283.

Yu X and Gabriel A. 1999. Patching broken chromosomes with extranuclear cellular DNA. Mol. Cell 4: 873-881.

Yuan JD, Shi JX, Meng GX, An LG, Hu GX. 1999. Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome. Cell Res. 9: 281-290.

Yuanxin Y, Chengcai A,Li L, Jiayu G, Guihong T, Zhangliang C. 2003. T-linker-specific ligation PCR (T-linker PCR): An advanced PCR technique for chromosome walking or for isolation of tagged DNA ends. Nucleic Acids Res. 31: e68.

Zardoya R & Meyer A. 1998. Cloning and characterization of a microsatellite in the mitochondrial control region of the African side-necked turtle, *Pelomedusa subrufa*. Gene. 216: 149-153.

Zhang DX and Hewitt GM. 1996a. Highly conserved nuclear copies of the mitochondrial control region in the desert locust *Schistocerca gregaria*: Some implications for population studies. Mol. Ecol. 5: 295-300.

Zhang DX and Hewitt GM. 1996b. Nuclear integrations: Challenges for mitochondrial DNA markers. Trends Ecol. Evol. 11: 247-251.

Zhang W, Bouffard GG and Wallace SS. 2007. Estimation of DNA sequence context-dependent mutation rates using primate genomic sequences. J. Mol. Evol. 65:207-214.

Zhao Z, Jin Li, Fu YX, Ramsay M, Jenkins T, Leskinen E, Pamilo P, Trexler M, Patthy L, Jorde LB, Ramos-Onsins S, Yui N, Lii WH. 2000. Worldwide DNA sequence variation in a 10-kilobase noncoding region on human chromosome 22. Proc. Natl. Acad. Sci. USA. 97: 11354-11358.

Zischler H, Geisert H and Castresana J. 1998. A hominoid-specific nuclear insertion of the mitochondrial D-loop: implications for reconstructing ancestral mitochondrial sequences. Mol. Biol. Evol. 15: 463-469.

Zischler H, Möss M, Handt O, von Haeseler A, van der Kuyl AC, Goudsmit J, Pääbo S. 1995a. Detecting dinosaur DNA. Science. 268: 1192- 193.

Zischler H, Geisert H, von Haeseler A, Pääbo S. 1995b. A nuclear fossil of the mitochondrial D-loop and the origin of modern humans. Nature. 378: 489-492.

Zouros E, Freeman KR, Ball AO, Pogson GH. 1992. Direct evidence for extensive paternal mitochondrial DNA inheritance in the marine mussel *Mytilus*. Nature. 359: 412-414.

Appendix A. List of primers and WGS trace files (TI) from Genbank used to determine the presence of target numts and generate their corresponding sequence. Table indicates the name, period of insertion (in million years, Ma) and location of individual numts in trace file reports.

Insertion Time	Locus Name	Taxon TI number (Position)	Primers
25-18 Ma	h2_181	N/A	F-GCAAAGGCCCTTCCTTCTT R-CTCCCACCTCCACCTCATT
	h4_179	N/A	F-TCCAAATTTCTCCTTTTGATAA R-CTTGGTCTGACTTGGGCAGT
	h9_367	*G. gorilla* AF035465	N/A
		N. leucogenys 1740323803 (68-383) 1893040019 (305-678) 2111902628 (30-129) 2055598196 (470-844) 2051768957 (607-854) 2071273899 (604-867) 2111242166 (836-978)	
	h4_316	*G. gorilla* 2036480803 (36-22) 2035770266 (447-130) 1666197786 (723-407)	N/A
		N. leucogenys 1903839244 (197-67) 2069579440 (262-22) 2100287284 (442-125) 2100976325 (462-145)	
	h4_3525	*G. gorilla* 1666197786 (406-36) 2035770266 (128-30)	N/A
		N. leucogenys 2100287284 (124-20) 2100976325 (144-25)	
	h7_75	*N. leucogenys* 2103608805 (513-444) 2083527588 (586-655) 1856862962 (219-288) 2055295397 (708-776) 2037471758 (86-155)	F-GCAGAAGCATCTAACAACAGG R-TCCTCCTGGAATTCAACCAT
18-14 Ma	2_171	*G. gorilla* 1677027367 (133-291) 2036454097 (660-814)	N/A
	h3_109	N/A	F2-CACTGGAGGAGGGTGATGATC R3-AGCACATTGGCTTTCCAGTAC
	11_86	N/A	F-AACTTGTTTGCTTTCAATGTCA R-GCAGCTGATGGGCTTTTTA
	h17_13321	*G. gorilla* 2019418209 (44-592)	N/A

		1670592741 (23-966) 1671775833 (25-752) 2033709513 (21-169)	
14-7 Ma	11_138	*G. gorilla* 1688527071 (767-937) 1674588214 (953-1124)	N/A
	h3_75	*G. gorilla* 2033326628 (794-868) 2033233491 (159-233)	N/A
	h5_3463	*P. troglodytes* ti268741600 (660-28)	N/A
		G. gorilla 2035181648 (559-68) 2018891626 (424-31) 1668869677 (386-38)	
	h10_113	N/A	F-CTGGGACAGTATTAATGCCA R-TTCAAATCTCAGTGTTGTGG
	h5_336	N/A	F-CTATCAACAGAACAGAATAC R-TCAATTCTTCGAAGTTGGAG
	11_2451	*G. gorilla* 1679297930 (429-23) 1679327622 (1052-27) 1666426381 (1145-489)	N/A
	h8_158	*G. gorilla* 1687938424 (216-372)	N/A
	h8_63	*G. gorilla* 1680524703 (293-355) 1680515464 (155-217)	N/A
7-6 Ma	h3_406	N/A	Fa-CTATCTATCTGAGAAAGGTC Ra-GAGATGTGTCTGTTCATGTC
	h6_185	N/A	F-CATAGCTGAACAAAAGGCAG R-GCAAATGTTGCTGCCTGATC
	5_2347	*P. troglodytes* 245056372 (413-69) 258203367 (600-763)	N/A

Appendix B. Insertion time and chromosomal location of Hominoidea numts derived from MCR, MT_P and MT_F. * Human assembly (March 2006). ** Chimpanzee assembly (March 2006). *** Orangutan assembly (July 2006).

Hominoidea:

Insertion time (Ma)	Locus name	Chromosomal location *
25-18	h2_181	chr2:56361105-56361285
	h4_179	chr4:5457167-5457345
	h9_367	chr9:34989142-34989510
	h4_316/h4_3525	chr4:65155015-65158855
	h7_75	chr7:110527934-110528008
	2_171	chr2:40865601-40865761
18-14	h3_109	chr3:68790791-68790899
	11_86	chr11:31533232-31533428
	h17_13321	chr17:21942648-21955968
14-7	11_138	chr11:110252926-110253096
	h3_75	chr7:110527934-110528008
	h5_3463	chr5:93928917-93932379
	h10_113	chr10:114644327-114644439
	h5_336	chr5:120394576-120394911
	11_2451	chr11:10486010-10488459
	h8_158	chr8:74060486-74060643
	h8_63	chr8:40047266-40047328
7-6	h3_406	chr3:43245822-43246227
	h6_185	chr6:125759417-125759601
	5_2347	chr5: 79981597-79983943
	2_158	chr2:227295229-227295386

Human:

Insertion time (Ma)	Locus name	Chromosomal location *
≤6	hY_146	chrY:19493376-19493521
	hY_77 1,2	chrY:22743283-22743359 chrY:22954129-22954205
	h2_132	chr2:149355765-149355896
	h4_131	chr4:55889084-55889214
	hX_284	chrX:125434116-125434399
	h13_256	chr13:108874473-108874728

Chimpanzee:

Insertion time (Ma)	Locus name	Chromosomal location **
≤6	pan6_105	chr6:10163792-10163896

	pan7_67	chr7:156987471-156987537
	pan18_64	chr18:69214911-69214974
	pan8_74	chr8:133603293-133603366
	pan8r_177	chr8_r:3603832-3604008
	pan9_1480	chr9:27823883-27825362
	pan8_294	chr8:340619-340892
	pan17_70	chr17:41418401-41418470
	pan3_570	chr3:82584001-82584570
	pan16_124	chr16:69963897-69964020
	panUn_818	chrUn:41177641-41178459
	pan8_1258	chr8:47844710-47848307
	pan7_1565	chr7:29312566-29314130
	pan1_75	chr1:179331902-179331976
	panY8000_1-16	chrY:14200750-14204541 chrY:9080566-9084356 chrY:9815553-9819340 chrY:14304854-14308648 chrY:8242882-8246675 chrY:9711430-9715223 chrY:13376620-13380412 chrY:4519417-4523211 chrY:8347010-8350802 chrY:8979779-8983571 chrY:863673-867464 chrY:13480626-13484416 chrUn:1417713-1421507 chrY:5276040-5279817 chrY:12618796-12622573 chrY:1620808-1624584

Orangutan:

Insertion time (Ma)	**Locus name**	**Chromosomal location *****
≤14	pgo14_177	chr14:32217963-32218139
	pgo3_1085	chr3:146875534-146876618
	pgo19_220	chr19:20973878-20974097
	pgo2b_446	chr2b:107002779-107003224
	pgo4_569(1)	chr4:194385369-194385966
	pgo4_104	chr4:71543737-71543840
	pgo8_110	chr8:1239714-1239823
	pgo18_135	chr18:41005234-41005368
	pgo10_70	chr10:121833553-121833622
	pgo8_273	chr8:40739917-40740208
	pgoX_78	chrX:77811121-77811198
	pgo5_172	chr5:114959974-114960233
	pgo16r_110	chr16_r:9950532-9950641

	pgo6_231	chr6:173022367-173022597
	pgo3_433	chr3:146022407-146022839
	pgo1_70	chr11:88325909-88325978
	pgo11_544(1)	chr11:88325444-88326232
	pgo8_78	chr8:18915312-18915389
	pgo16_166	chr16:24806968-24807133
	pgo6_445	chr6_r:9944899-9945337
	pgo19r_126	chr19r:2162455-2162580
	pgo2a_182	chr2a:55952847-55953028
	pgo10_180	chr10:6713458-6713639
	pgo19_280	chr19:14698462-14698741
	pgo3_71	chr3:115571762-115571832
	pgo3_77	chr3:53115723-53115799
	pgo4_937	chr4_r:11865365-11866301

Appendix C. Name, chromosomal location and age in million years (Ma) of each hominoid numt. Taxon-specific numts are shown: Human-specific (H.S.), chimpanzee-specific (C.S) and orangutan-specific (O.S).

Insertion time (Ma)		Name	Location in human Mt (blt36)	Nuclear Location (March 2006)
24-18	Cluster 1	h4_236	9170-9405	4:65154693-65154928
		h4_60	9168-9109	4:65154963-65155022
		h4_316	115-428	4:65155015-65155330
		h4_3525	13037-16568	4:65155331-65158855
		h4_1345	9467-10837	4:65158856-65160200
	Cluster 2	8_68(1)	12960-13064	8:49475800-49475904
		8_68(2)	13941-14053	8:49475906-49476018
		h9_367	16102-16471	9:34989151-34989507
		h7_75	359-425	7:110527934-110528008
		h2_181	16131-16302	2:56361105-56361282
		h4_179	16297-16474	4:5457167-5457345
		5_503	3823-4316	5:60093116-60093608
		h4_616	4614-5234	4:14116587-14117202
		2_1107	4855-6208	2:155828212-155829566
		h8_386	5228-5613	8:70177762-70178147
		11_204	13256-13453	11:47302111-47302308
		1_211	2468-2677	1:5832905-5833115
		6_112	8490-8693	6:1651211-1651414
		5_50	2005-2060	5:118490993-118491048
18-14	Cluster 3	h17_13321	14366-16571-11113	17:21942648-21955968
		17_232	14328-14583	17:21955965-21956219
	Cluster 4	h3_109	16087-16193	3:68790791-68790899
		3_76	4375-4430	3:68790897-68790972
		3_136	3060-3217	3:68790970-68791128
		11_86	16199-16389	11:31533232-31533425
		h4_152	2901-3052	4:27341144-27341295
		3_58	4706-4768	3:123890271-123890346
		8_99	6883-7010	8:121305733-121305860
		2_592	6271-6999	2:50669330-50670057
		2_171	958-1126	2:40865601-40865761
		4-278	2005-2390	4:129222010-129222387
14-7		h5_336	380-710	5:120394576-120394911
		h10_113	348-461	10:114644327-114644439
		h5_3463	12663-16125	5:93928917-93932379
		11_2451	523-2974	11:10486010-10488459
		12_68	4242-4309	12:61454057-61454124
		11-138	15525-15612	11:110252926-110253096
		h3-75	15561-15635	3:63807742-63807816
		h8_158	627-784	8:74060486-74060643
		h8_63	808-870	8:40047266-40047328
7-6		5_2347	343-2699	5:79981597-79983943
		h3_406	15810-16213	3:43245822-43246227
		h6_185	16104-16284	6:125759417-125759601

		2_158	892-1050	2:227295229-227295386
H.S.	Cluster 5	hX_749 hX_284 hX_554	6554-7303 971-686 10607-11160	X:125433368-125434116 X:125434116-125434399 X:125434395-125434948
		hY-77_1	557-629	Y:22954129-22954205
		hY-77_2	557-629	Y:22743283-22743359
		h13_256	984-1239	13:108874473-108874728
		h2_132	613-744	2:149355765-149355896
		h4_131	964-1094	4:55889084-55889214
		hY_146	15567-15712	Y:19493376-19493521
C.S.	Cluster 6	pan6_85 pan6_105	11463-11379 15491-15595	6:10163706-10163790 6:10163792-10163896
		pan16_124	16103-16227	16:69963897-69964020
		pan17_70	16025-16094	17:41418401-41418470
		panUn_818	15414-16227	Un:41177641-41178459
		pan7_1565	16355-16571; 1-1358	7:29312566-29314130
		pan3_570	15569-16132	3:82584001-82584570
		pan8_1258	6272-8442; 15722-405	8:47844710-47848307
		pan8_294	15769-16061	8:340599-340892
		pan8r_177	15504-15679	8_r:3603832-3604008
		pan9_1480	14242-15748	9:27823883-27825362
		pan8_74	15611-15684	8:133603293-133603366
		pan1_75	889-962	1:179331902-179331976
		pan18_64	15936-15999	18:69214911-69214974
		pan7_67	15681-15747	7:156987471-156987537

O.S.	Cluster 7	pgo4_569(1)
		pgo4_569(2)
			15317-15885
			3848-3879
				4:194385398-194385966
				4:194385369-194385442

	Cluster 8	pgo11_544 (1) pgo11_544 (2)	7270-7517 16100-1-71	11:88325975-88326232 11:88325444-88325988
		pgo1_70	16563-1-60	1:72964446-72964515
		pgo10_180	104-390	10:6713458-6713639
		pgo10_70	16083-16154	10:121833553-121833622
		pgo14_177	15532-15711	14:32217963-32218139
		pgo16_166	16555-1-150	16:24806968-24807133
		pgo16r_110	16383-16492	16_r:9950032-9951141
		pgo18_135	16010-16144	18:41005234-41005368
		pgo19_220	15508-15728	19:20973878-20974097
		pgo19_280	8-390	19:14698462-14698741
		pgo19r_126	162-390	19_r:2162455-2162580
		pgo2a_182	104-390	2a:55952847-55953028
		pgo2b_446	15440-15885	2b:107002779-107003224
		pgo3_1085	14640-15724	3:146875534-146876618
		pgo3_433	16105-16538	3:146022407-146022839
		pgo3_71	281-451	3:115571762-115571832
		pgo3_77	587-662	3:53115723-53115799
		pgo4_104	15909-16011	4:71543737-71543840
		pgo4_937	16393-1-892	4_r:11865365-11866301
		pgo5_172	16036-16291	5:114960062-114960233
		pgo6_231	16210-16500	6:173022367-173022597

		pgo6_445	16378-1-244	6:9944899-9945337
		pgo8_110	15934-16037	8:1239714-1239823
		pgo8_273	15940-16230	8:40739917-40740208
		pgo8_78	74-151	8:18915312-18915389
		pgoX_78	16165-16245	X:77811121-77811198

Appendix D. List of primers used for amplification and sequencing of target numts. Also, trace file and location of numts filtered from the Genebank. Reference sequences in the gibbon genome were taken from trace files in *Nomascus leucogenys* and direct sequencing was done in *Hylobates lar*.

* Primer 2 in Figure 1 of Zischler et al. (1998).

Insertion time (Ma)		Name	Taxon	Trace file Acc. N.	Trace position	Primers
24-18	Cluster 1	h4_236	*Gorilla*	2036480803	358-123	N/A
				2035770266	769-534	
			Nomascus	1903839244	526-289	
				2069579440	589-354	
				2100287284	760-534	
				2100976325	789-554	
		h4_60	*Gorilla*	2036480803	88-29	
				2035770266	499-440	
			Nomascus	1903839244	249-190	
				2069579440	314-255	
				2100287284	494-435	
				2100976325	514-455	
		h4_316	*Gorilla*	2036480803	36-22	
				2035770266	447-130	
				1666197786	723-407	
			Nomascus	1903839244	197-67	
				2069579440	262-22	
				2100287284	442-125	
				2100976325	462-145	
		h4_3525	*Gorilla*	1666197786	406-36	
				2035770266	128-30	
			Nomascus	2100287284	124-20	
				2100976325	144-25	
		h4_1345	*Gorilla*	1218235913	146-645	
			Nomascus	1749374728	67-339	
				1722101939	86-357	
	Cluster 2	8_68-1	N/A			F-ccactgagaaaggagacagc
		8_68-2				R-aacatgaggaaaattcagggt
		h9_367				Zi98F-gcgagctgaaggactttctg (Gibbon) * or F- ctcccgagtagctgggattac (Gorilla)

						R- ggatttgcagctgtgttca
		h7_75	*Nomascus*	2103608805 2083527588 1856862962 2055295397 2037471758	513-444 586-655 219-288 708-776 86-155	F-gcagaagcatctaacaacagg R-tcctcctggaattcaaccat
		h2_181	N/A			F-gcaaaggcccttccttctt R-ctcccacctccacctcatt
		h4_179	N/A			F-tccaaatttctccttttgataa R-cttggtctgacttgggcagt
		5_503	N/A			F-gggaacagcttttgttgct R-catgcattggcacttctgt
		h4_616	*Nomascus*	1891975446 1813822831 2094418442 2062918683 2038394904 2110934940 2043174271 1746341073	154-20 623-752 334-23 460-25 85-702 926-491 24-415 810-536	F-tgttgttagctggttgctatgc Ra-aatacccagcctactcctcctc Fa-trggatcaggggtgttaatc R-tccagccaaactaagcttcata 2 overlapped fragments
		2_1107	N/A			Fa-agtccttaggtattgcaga R4-ctatgttcctcatgttttag
		h8_386	N/A			F-tatcaaaggcccagaaggag R-ggttttagatgagaaatctgttgtc
		11_204	*Nomascus*	2037611582 2036714198 2105643423	117-313 590-394 492-688	F-tctgagacccagctcaaca R-gtttgatgctttgctgtcg
		1_211	N/A			F-agccagggtagaggcaag R-ctggggacagagctcactt
		6_112	N/A			F-gacccacacagcaatgaga R-tcccttgactcccctgtt
		5_50	N/A			F-ttcaacatcagcaacaccc Ra-acatgacgaaaccccatctc
18-14	Cluster 3	h17_13321 (530bp)	*G. gorilla*	2019418209 1670592741 1671775833 2033709513	44-592 23-966 25-752 21-169	N/A
		17_232	*G. gorilla*	2019418209 1670592741	589-844 963-1210	Fa-ctgtctttctcctctgaccc Ra-aatgctaagtttcccagtag

				1671775833 2033709513	749-1002 166-420	Obtained through both direct sequencing and Trace files. Trace files are optional here
	Cluster 4	h3_109	N/A			F2-cactggaggagggtgatgatc R3-agcacattggctttccagtac
		3_76				
		3_136				
		11_86	N/A			F-aacttgtttgctttcaatgtca R-gcagctgatgggcttttta
		h4_152	N/A			F-tggactcacagtttcacatgg R-atttaggggacagggttaca
		3_58	N/A			F-ccctaaggctgggactctt R-tttctctgctcttgccctt
		8_99	N/A			F-tagaccctgccctcatctc R-tggggcatataaaagatgaaa
		2_592	N/A			Fa-attggctttttggagtttac Ra-ctgatccgtcctaatcacag Fb-ctatgcccatatacccgaat Rb-gtcctagctactcaggaggc 2 overlapped fragments
		2-171	*G. gorilla*	1677027367 2036454097	131-291 658-814	N/A
		4-278	N/A			F-gggcttctttgtgtcaagg R-gggatccaggtctttcttg
14-7		h5_336	N/A			F-ctatcaacagaacagaatac R-tcaattcttcgaagttggag
		h10_113	N/A			F-ctgggacagtattaatgcca R-ttcaaatctcagtgttgtgg
		h5_3463	*P. troglodytes*	268741600	660-28	N/A
			G. gorilla	2035181648 2018891626 1668869677	559-68 424-31 386-38	
		11_2451	*G. gorilla*	1679297930 1679327622 1666426381	429-23 1052-27 1145-489	N/A
		12_68	N/A			F-catctatgcctaacctaaca R-ttcactctctgagccggttt
		11-138	*G. gorilla*	1688527071	767-937	N/A

		h3_75	*G. gorilla*	2033326628 2033233491	794-868 159-233	N/A
		h8_158	*G. gorilla*	1687938424	215-372	N/A
		h8_63	*G. gorilla*	1680524703 1680515464	293-355 155-217	N/A
7-6		5_2347	*P. troglodytes*	245056372 258203367	413-69 600-763	N/A
		h3_406	N/A			Fa-ctatctatctgagaaaggtc Ra-gagatgtgtctgttcatgtc
		h6_185	N/A			F-catagctgaacaaaaggcag R-gcaaatgttgctgcctgatc

Appendix E. Origin and accession number of human whole mitochondrial genomes of world populations (Mitochondrial Genome Database; Ingman & Gyllensten 2006).

Africa (n=20) DQ112692, DQ112699, DQ112707, DQ112714, DQ112721, DQ112728, DQ112742, DQ112756, DQ112792, DQ112794, DQ112796, DQ112850, DQ112857, DQ112883, DQ112904, DQ112911, DQ112919, DQ112933, DQ112960, DQ112961.
Asia (n=20) DQ112779, DQ112781, DQ112784, DQ112786, DQ112788, DQ112859, DQ112864, DQ112866, DQ112868, DQ112873, DQ112875, DQ112878, DQ112880, DQ112882, DQ112928, DQ112930, DQ112935, DQ112939, DQ112951, DQ112954.
Europe (n=20) DQ112760, DQ112764, DQ112769, DQ112795, DQ112805, DQ112809, DQ112813, DQ112828, DQ112831, DQ112836, DQ112837, DQ112841, DQ112842, DQ112891, DQ112936, DQ112941, DQ112942, DQ112943, DQ112945, DQ112955.
India (n=10) DQ246811, DQ246813, DQ246815, DQ246817, DQ246819, DQ246822, DQ246824, DQ246826, DQ246828, DQ246831.
Australia (n=10) DQ112750, DQ112751, DQ112752, DQ112753, DQ112754, DQ404441, DQ404443, DQ404445, DQ404446, DQ404447.
North America (n=5) DQ112846, DQ112870, DQ112872, DQ112888, DQ112889.
South America (n=5) DQ112772, DQ112774, DQ112776, DQ112832, DQ112871.
Jewish (n=5) DQ301789, DQ301795, DQ301805, DQ301811, DQ301812.
Melanesia (n=5) DQ112886, DQ112887, DQ112895, DQ112896, DQ112897.

Appendix F. Approval from the Association of Zoos and Aquariums of the USA.

GORILLA
SPECIES SURVIVAL PLAN®

ASSOCIATION
OF ZOOS &
AQUARIUMS

Nicola Mary Anthony
Department of Biological Sciences
University of New Orleans
New Orleans LA 70148

October 16, 2007

Dear Drs. Anthony and Jensen-Seaman,

The Gorilla SSP® has reviewed your proposal and voted to APPROVE your request for gorilla genetic samples. SSP approval provides confirmation that the management group and advisors have reviewed the proposal and found it to be of sound scientific merit. We encourage institutions to participate if they are able to do so. We ask you to note that you will still need to contact individual zoos with your proposal and requests.

We wish you the best in your research endeavor and look forward to seeing results when they are available.

Sincerely,

Kristen E Lukas

Kristen E. Lukas, Ph.D.
Chair, Gorilla Species Survival Plan

Curator of Conservation and Science
Cleveland Metroparks Zoo
3900 Wildlife Way
Cleveland, OH 44109
P: 216-635-2523
F: 216-635-3318
E: kel@clevelandmetroparks.com

Appendix G. Presence/absence polymorphism of three numts in 68 gorillas captive in USA zoos. The 18 underlined names represent wild-born gorillas and one captive-born gorilla (Kwanza) for which all known ancestors exhibit the same mitochondrial haplogroup. These gorillas were used to assess mtDNA haplogroup differences in the distribution of polymorphic numts (see Table 3.3).

House Name	Studbook *	Mitochondrial Haplogroup	Gcl18_1	Numt1_1	Numt2_1
Paki	191	C1	+/+	-	+/-
Banga	224	C1	+/-	-	+/-
Stadi	1186	C1	+/+	-	+/-
Charlie	1409	C1	+/+	-	+/-
Casey II	801	C1	+/-	-	+/-
Ramar	537	C1	+/-		+/+
Ntondo	1301	C1	+/-		+/-
Abe	52	C1	+/-		+/+
Motuba/Tubby	883	C1	-/-		-/-
Kitombe/Ma	934	C1		-	
Rok	701	C1		-	
Willie B II	115	C1	+/-	+	+/-
Kinyani	820	C1	+/-	+	+/-
Kekla	1108	C1	+/-	+	+/-
Mbeli	1693	C1	+/-	-	-/-
Chicory	890	C1	-/-	+	+/-
Mosuba	835	C1	-/-	-	-/-
Taz	1110	C2	+/-	-	+/+
Ozoum	175	C2	+/-	-	-/-
Donna	336	C2	+/+	-	-/-
Shango	1123	C2	-/-		-/-
Shamba	221	C2	+/-	-	-/-
Machi	609	C3	-/-	-	-/-
Mia Moja	1109	C3	+/-	-	-/-
Kashata	1294	C3	+/-	-	-/-
Holoki	393	C3	-/-	-	-/-
Choomba	180	C3	-/-	+	-/-
Kudzoo	1330	C3	+/-	+	+/-
Olympia	1410	C3	-/-	+	+/-
Kidogo	1484	C3	-/-	+	+/-
Sukari	1485	C3	+/-	+	+/-
Mumbah	379	C3	-/-	-	+/-
Chaka	864	D1	+/-	+	+/-
Ivan	710	D2	+/-	-	-/-
Beta	160	D2	+/+	-	-/-

Aqualina	781	D2	+/-	-	-/-
Praline	1418	D2	+/-	-	-/-
Fredrika	528	D2		-	+/-
Binti	556	D2		-	-/-
Alpha	159	D2			-/-
Curtis	1331	D2	+/-	+	+/-
Carlos	506	D3	+/-	-	-/-
Tabibu	1264	D3	+/-	-	-/-
Rollie	1414	D3	+/-	+	-/-
Bebac	872	D3	-/-	+	-/-
Jasiri	1486	D3	+/-	-	+/-
Kwanza	1107	D3	+/-	-	+/-
Susie	T1193/1835	D3	+/-	+	+/-
Katie	498	D3	+/+	-	+/-
Toni	432	D3	-/-	-	+/-
Jimmy Jr.	716	D3	+/-	-	+/-
Katoomba	168	D3	-/-	-	+/+
Bahati	1142	D3	+/-	+	+/+
Josephine	524	D3		-	+/-
Bombom	612	D3	-/-	-	+/-
Chipua	1419	D3	+/-	+	-/-
Kowali	663	D3	-/-	-	+/-
Makari	949	D3	+/-	+	-/-
Bulera	1120	D3	-/-	+	+/-
Madini	1413	D3		+	+/-
Azizi	1750	D3	+/-	+	+/-
Mokolo	948	D3		+	+/-
Kubandu	812	D3	-/-	-	+/-
Billy	1148	D3	+/-	+	+/-
Kimya	1345	D3	+/-	+	+/-
Sunshine	509	To be assigned		-	+/-
Koola	1369	To be assigned	+/+	-	+/+
Frank	265	To be assigned		+	

* Wharton (2007).

Appendix H. Sequence of 22 hits for putative gorilla numts detailing the locus name and corresponding coordinates in the human genome (hg18).

```
AAGL1682        (chr19:17790170)
CTGGGATTGTGGGGGCAATGAATGAAGCGAATAAATTTTCGTTCATTTTGGTTCTCAGGGTTTACAGAAGTTTTTTATTTTTATAGTTTTTGGTAAG
GGG

AAGL1145        (chr1:113276030)
ATGGCCATCGCTGTAGTATACCCAAAAACAACCATCATCCCCCCTAATAAATTAAAAAAACCATTAACCCATATAACCCTCCCCACAAGTTTAAAAT
ATAGCCCACCCCAACCACACCACTACNNNNN

CABD5746        (chr9:23156894)
GTGCATAAGTAGGTGACCTGCAGTGATAGTAGCGGTTCCCCTCCAAAGGGGAACGTGTGGGCAGTTTTAGATGTTATGGCCCTGAAGCAAGAACCAG
ATGCCGGATACTGTTCATTCTAGCTACCCACAAGTGTTATGGGCCCGGAGCGAGGAGAGTAGCACTCTTGTGCGGGATATTGATTTCACGGAGGATG
GTGACTAAGGGACTCCTATCTGAGGGGGGGCATCCGTGGGGGCAAGAAAGGATTTGATTGTAATGTGCTATGTACGATAAATGATTGTATGTGCYAT
GTACTGTCGAGGATGGACAGGTCTGTTGATATTCTAAGGGTTGGGGATTGTCCTTGGAGGTAGGGTTAATGTTCGATAGTTGTGAGGGTCGATCGCT
GTACGTGCTTGTAAGCATTGGGAGGAGGTTTTAATGTGGGATGGGTTCTGTATGTACTA

Go11_188        (chr11:87805752)
TTAGTTCTTCTGTAAGGTAATAGATTGGTCCAATTGGGTGCAAGTAGTTCAGTTGTATGTTTGGGATTTTTTAGATAATAGGTGTCGAGCTTGAACG
CTTTCTTAATTGGTGGCTGCTTTTAGGCCTACTATGGGTATTAAATTTTTTACTCTCTTTACAAGGTTTTTTCCTNGTGTCCAAAGAGCTGTTCCTN
TTTGGACTAACAGTTAAATTTACAGGGGTTTTGNAGGGTTCTGNGGGGAANNTTAAAGTNGANCTAAGANTCTANCTTGGNCAACCAGCTNTCACCA
GNCTCGGTAGGNTNNTCGCCNCTNNCTGNNNNTCTNCCCACTATTTTGCTACNTANACGGGNGTNCTCTTTTAGCTNNNCTNAGGAAGCTCNTNTGG
NTNCNGGGGGNTTAGCTTTNGTTNTNTTTGCAAAGTTATTTCTNGTTANTTCANNNTGCAGNAGNNACNAGGTNTNGNCCTTGCTGTATTGTGCNNG
GTTATNATTTTTCANCTTTCCCTTGCGGTNCTNTATCTNTNGCGCCANATTACAATTTCTATCNNCTATACTTTNTTTGAGTANATGGTTTGNTTAN
AGTTGTNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNTTNCTTTGNGCCTTCGTCAGGGTTTGNTGAAGATGGCGGTATATAGGCTGAGCAAGAGGTGGTG
AGGTTGATCGGGGTTTATCGATTACAGAACAGGCTCCTCTAGAGGGATNTGAAGCACCGCCAGNGTCCTTTGAGTTTTAAGCTGTGGCTCGTAGTGT
TCTGGCGAGCAGTTTTGTTAATTTAACTGTTGAAGTTTAGGGCTAAGCATAGTGGGGTATCTAATCCCAGTTTGAGTCTTAGCTATTGTGTGTTCAG
ATGCGTTAAAGCCACTTTCGTAGTTTATTTTGTATCAACTGGAGTTTTTTACAACTCAGGTGAATTTTAGCTTTATTGAGGGAAATTGATCTAAAAC
GCTCTTTACGCCGGTTTCTATTGACTTGGGTTAATCGTGTGACCGCGGTGGCTGGCACGAAATTGACCAACCCTGGGGTTAGTATAGCTTAGTTTAA
CTTTCGTTCATTGCTAAAGGTTTATCACTGCTGTCTCCCGTGGGGGTGTGGCTAGGCTAAGCGTTTTGAGCTGCATTGTTGTGTGCTTGATACCTGT
TCCTTTTGATCGTGGTGATTTAGAGGGTGACTCACCGGGGCGGGGTTGCTTGCATGTGTAATCTTACTAAGAGCTAATAGAAAGGCTAGGACCAAAC
CTATTTGTTTATGGGGNGGTATGAGCCCGTCGAAACATTTTCAGTGTATTGCTTTGAGGAAGTAAGCTACATAAACCGTATGGGGTGTCTTTGGGGT
TTGGTTGGTTCGGGGTATGGGGTTAGCAGCGGTGTATATGTTGAGTAAGGTGGGTAGGAGTTGCATTGGCAGGGTTAGTAGGATGGGAGTTGAGGGA
GGAGAATATGTTAGTTGAGGGGTGACTGTTAAAAATGCATACCGCCAAAAGATGAAAAATCCGGTTAGGCTGATGTTAGGGCTCTTTGTTTTTGGGG
TTTGGCAGAGATGTTTTTGAGTGCTATGGCCAGAGGTGGGGGGAGGGGGAGGGTTGTGGAAATTTTTATTGTAGTATTGGTGTGAAGAGCGGTTGCG
TGCGCATTCGTTGGCTATTGCTATGTCCAACAAGCATGGATTAATTAACACATTATGGTAGTTATGCTCGCCTGTAATATTGAACGTAAGTGCGATA
AATAATGGGATGGGGCAGGAATCAAAGACAGATACTGCGACATAGGGTGCTCCGGGGCCAGCGTTTCGCAATGCTATCGCGTGCACACCCCCCAGAC
GAAAATACCAAATGCATGGAGAGCTCCCGTGACTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAACGTGTGGGCG
ACTTTGGGTGTTATGGCCCTCAAGTAAGAACCAGATGCCGGATACAGTTCATTCTAGCTACCCCCAAGTGTTATGGGCCCGGAGCGAGGAGAGTAGC
A

Numt1_1         (chr1:224874831)
GTGCGTTAATTAATAACCATGAATTAATTAACACCATGAAGCATATTGCGCTCGGCTGTAATATTGAACGTAGGTGCGATAAATAATGGTATGGGGT
AGGAATCAAAGACAGATACTGCGACATAGGGTGCTCCGGTTCCAGCGTTTCGCAATGCTATTGCGTGCACGCCCCCCCCCGACGAAAATACCAGATG
CATGGAGAGCTCCCGTGAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAACGTGTGAAGCGCTTTAGGTGTTAT
GACCCTGAAGTAGGAACCAGATGTCGGATACAGTTCACTTTTAGCTACCCCCAAGTGTTATGGGCCCGGAGCGAGGAGAGTAGCACTCTTGTGCGGG
ATATTGATTTCACGGAGGATGGTGTTCAAGGGACCCCCATCTGAGGGGGGGCATCCATGGGGGCGAGGACGATTTAACTGGAATGTGCTATGTACGG
TAAATGATTTTATGTGTTATGTACTTTTGTAGAGGGTAGGTCGGTTGATATTTCGTTGGGTAGAGCAGGGGATGGGGGGGGTTTGTATGTGTTATAG
GTATTTGGGTGTTTATAGTACTGTATATTATTCATGGTGACTGGCAGTAATGCACGACATACATAACAATTATTGGTGGGTTAGCTAATACTTGGGT
GGTACCCAAATTTGTCTCCCCATGAAAGAACAGAGAATAGTTTAAATTAGAATGTTAGCTTTGGGTGCTGATGGTGGAGTCGAGGACTTTTTCTCTG
AATATGCCCTTGGAAGGAGGTCTTCGTTTCCRGTTYACAAGACTGGTGTATTGGTCTGTACTACAAGGGCAGGTTCATTTGAGTATTTTGTTTTCGA
TTAGGGATGTGACTGGCATTAGGAATAGGATTGTCGTGAAGTATAGTACGGATGCTACTTGCCCAATGGTGATGAAGGGGTAGCTTACTGGTTGTCC
TCCGATTCAGGTTAGAGTGAGGAGGTCTGTGATTAGGAATCAGTAGAGTAGTTGGCTTAATGGGCGGAATATTATGCTTTGTTGTTTGGACATGTGG
AGAACAGGAATTATTGCTAGGATGAGAATAGATAGTAATAGGGCTAAGACGCCTCCTAATTTATTGGGGACAGATCGGAGAATTGCGTAGGCAAATA
GGAAGTATCATTCGGGTTTGATGTGGGGTGGGGTGTTTAGGGGGTTGGCTAAGGTGTAGTTGTCTGGGTCTCCTAGGAGGTCTGGTGAGAATAGTGT
TAATGCTATTAGGGTCAGGAGAAAGAGGAATAGGCCTAGGATGTCTTTGATTGTGTAGTAGGGGTGGAAGGTAATTTTGTCAGAGTGGGAGTTAATT
AATTATTAATTAATTAACTAACTAATTAATTAATTAATTAACATGC

Numt2_1         (chr2:155454827)
CCCACTGGGGCGGGGATGCTTGCATGTGTAATCCTACTAAGAGTTAATAGAAAGGCTAGGACCAAACCTATTTGTTTATGGGGTGATGTGAGCCCGT
CGAAACATTTTCAGTGTATTGCTTTGGGGAGGTAAGCTACATAAACTGTGTGGGGTGTCTTTGGGGTTTGGTTAATTCGGGGTATAGGATCAGCAGC
AGTGTGTGTTGCTAGGGCGGGTKGGGGTTGTACTGGTGGGGTTGGTGGGGTGGGTGCTATGTTAGTTGAGGGGTGACTGTTAAAAATGCATACCGCC
AAAAGATGAAATTTGGAGTTTGGTAGGGCTGTTTTCTAGTGCTGAGGGGGTTTGGATTTTTTTGTTGTGTTTTTTGGTGTGAAGGGTGGTTGTGTTC
ACGTCCACCTGGTTGTTTTATGTCCAACAAGCATGAATTAATTAACACCATAAGGCATATTGCGCTCGGCTGTAATATTGAACGTAGGTGCGATAAA
TAATGGTATGGGGCAGGAATCAAAGACAGATACTGCGACATGTGGTGCTCCGGCTCCAGCGTTTCGCAATGCTATCGCGTGCACACCCCCCGACGAA
```

```
AAATACCAAATGCATGGAGAGCTCCCGTGAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAACGTGTGAAGCGC
TTTAGGTGTTATGACCCTGAAGTAGGAACCAGATGCCGGATACAGTTCATTCTTAGCTACCCCCAAGTGTTATGGGCCCGGAGCGAGGAGAGTAGCA
CTCTTGTGCGGGATATTGATTTCACGGAGGATGGTGTTCAAGGGATTCCTATCTGAGGGGGGGCATCCGTGGGGGCGAGGATGATTTAACTGGAATG
TGCTATGTACGATAAATGACTTTATGTGTTATGTACTTTTTGTGAGGGATGGATTGGTTGGTATTCCGTTGGGTGGAGCAGTGAGGGGGGGGGGTTG
TATGTGTTACAGGTGGTTGGACATTTGTAGTACTGTGCATTATTCATGGTGGCTGGCAGTAATGCACGACATACATGACAATTATTGATGGGTTAGC
TAATACTTGGGTGGTACCCAAATTTGTCTCCCCATGAAAGAACAGAAGAATAGTTTAAATTAGAATGTTAGCTTTGGGTGCTGATGGTGGAGTCGAG
GACTTTTTCTCTGAATATGCCCTTGGAAGGAGGTCTTCGTTTCCGGCTTACAAGACTGGTGTATTGGTCTGTACTACAAGGGCAGGTTCATTTGAGT
ATTTTGTTTTCGATCAGGGATGTGACTGGTATCAGGAATAGGATTGTCGTGAAGTATAGTACGGATGCTACTTGCCCAATGGTAATGAAGGGGTAGC
TTACTGGTTGTCCTCCGATTCAGGTTAGGGTGAAGAGGTCTGCGATTAGAAATCAGTAGAGTAGTTGGCTTAATGGGCGGAATATTATGCTTTGTTG
TTTGGATATGTGGAGAATAGGAATTATTGCTAGGATGAGAATAGATAGTAATAGGGCTAAGACGCCTCCTAGTTTATTGGGGACAGATCGGAGAATT
GCGTAGGCAAATAGGAAGTATCATTCGGGTTTGATGTGGGGTGGGGTGTTTAGGGGGTTGGCTAAAGTGTAGTTGTCTGGGTCTCCTAGGAGGTCTG
GTGAGAATAGTGTTAATGTTATCAGGGTCAGGAGAAAGAGGAATAGGCCTAGGATGTCTTTGATTGTGTAGTAGGGGTGGAAGGTGATTTTGTCAGA
GTGGGAGGGGATGCCTAGAGGGTTGTTTGATCCTGTTTCGTGTAGAAATAGGAGATGGAGGGTTGTTAGGGCTGTGATAATGAAGGGTAGGATAAAG
TGGAAGGTAAAGAATCGTGTAAGGGTAGGGCTATCTACTGAGTAACCACCTCAAACTCATTGGACTAGGTCTGTTCCGATGTATGGGATGGCGGATA
GCAAGTTTGTGATTACTGTGGCTCCTCAGAAGGATATTTGGCCTCATGGGAGGACATAGCCTATGAAGGCTGCTGCTATGGTTGTGAGTAGGAGGAT
GATGCCGATGTTTCAGGTTTCTTGGTAGAGAAATGAGCCGTAGTATAGGCCTCGGCCGATGTGTAGAAAGAGGCAAATGAAGAATATTGAGGCGCCG
TTAGCGTGGAGGTAGCGGATGGTTCAGCCATAGTTTACATCTCGGGTGATGTGAGCGATTGATGAGAAGGCGGTTGAGGCGTCAGGTGAGTAGTGTA
TGGCTAGGAATAGCCCTGTGGTGATTTGAAGGATTAAGCAGGTACCAAGGAGTGAGCCGAAGTTTCATCATGTGGAGATGTTGGACGGGGTAGGGAG
GTCAATGAATGAGTGGTTAATTAGTTTTGCTAGTGGGTTAGTTTTGCGTATAGGGGTCATTGATGTTCTTGTAGTTGAAGTACAACGATGGTTTTTC
ATATCATTGGTCGTGGTCGTGGTCCGTGCGAGAATGATGACATATGTTTTATTTTTATTGAGTGTGGGTTTAGT

Gcl18-1          (chr5:123205257)
CCCCCCGACGAAGATACCAGATGCATGGAGAGCTCCCGTGAGTGGTTAATAGGGTGATAGACCTGTGATCCATCGTGATGTCTTATTTAAGGGGAAC
GTGTGAAGCGCTTTAGGTGTTATGACCCTGAAGTAGGAACCAGATGTCGGATACAGTTCATTTTTAGCTACCCCAAGGTGTTATGGGCCCGGAGCGA
GGAGAGTAGCACTCTTGTGCGGGATATTGATTTCACGGAGGATGGTGTTCGAGGGATCCCCATCTGAGGGGGGGCATCCGTGGGGACGAGGGTAATT
TAACTGGAATGTGCTATGTACGATAAATGACTTTATGTGTTATGTACTTTTGTAGGAGGCAGGTCGGTTGATATTTCGTTGGGTGGAGCAGGAGATG
GGGGGGGTTGTATGTGTTACAGGTGTTTGGGTGTTTATAGTACTGTGCATTATTCATGGTGACTGGCANNNNNN

Gcl18-2          (chr13:113308247)
AAGTACATAACACATAAGATCATTTATCGTACATAGCACANNNNNN

Go1_308          (chr1:183761643)
CCCTCAACTCCCATCCTACTAACCCTGCCAATGCAACTCCTACCCACCTTACTCAACATATACACCGCTGCTAACCCCATACCCCGAACCAACCAAA
CCCCANAGACACCCCATACGGTTTATGTAGCTTACTTCCTCAAAGCAATACACTGAAAATGTTTCGACGGGCTCATACCGCCCCATAAACAAATAGG
TTTGGTCCTAGCCTTTCTATTAGCTCTTAGTAAGATACACATGCNAGCAACCCCCGTCCCGGTGAGTCACCCCTCTAATCACCACGATCANAAAGGA
CACGTATCAAGCACGCANNNNNN

Go2_201          (chr2:33645118)
GTTGCCCCTTTGTTGTTCAAGTGTTTGTAGTACTGTATATTATTCATGGTGACTGGCAGTAATGCCCGATATACATGGCGGTTATTGGTGGGTTAGC
CAATACTTGGGTGGTACCCAAATTTGCTTCCCCATGAAAGAACAGAGAATAGTTTAAATTAGAATGCTAGCTTTGGGTGCTGATGGTGAAGTCAAGA
ACTTTTTCTCTGATTTGCCCTGGAAGGAGGNNNNNN

Go2_437          (chr2:101487391)
CCGTACATAGCACATTCCAGTTAAATCGTCCTCGCCCCCATGGATGCCCCCCCTCAGATGGGGGTCCCTTGAACACCATCCTCCGTGAAATCAATAT
CCCGCACAAGAGTGCTACTCTCCTCGCTCCGGGCCCATAACACTTGGGGGTAGCTAAAAGTGAACTGTATCCGACATCTGGTTCCTACTTCAGGGTC
ATAACACCTAAAGCGCTTCACACGTTCCCCTTAAATAAGACATCACGATGGATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATG
CATCTGGTATTTTCGTCGGGGGGGGGCGTGCACGCAATAGCATTGCGAAACGCTGGAACCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCT
ACCCCATACCATTATTTATCGCACCTACGTTCAATATTACAGCCGAGCGCGAATCCACCACAGTGCGAGATCATCGGACCNNNNNN

Go3_528          (chr3:129446363)
AGAGCTAAGACGCCTCCTAGTTTATTGGGGACAGATCGGAGAATTGCGTAGGCAAATAGGAAATATCATTCGGGTTTGATGTGGGGTGGGGTGCTCA
GGGGGTTGGCTAAGGTGTAGTTGTCTGGGTCTCCTAGGAGGTCTGGTGAGAATAGTGTTAATGTTATCAAGGTCAGGAGAAAGAGGAGTAGGCCTAG
GATGTCTTTGATTGTGTAGTAGGGGTGGAAGGTGATTTTGTCAGAGTGGGAGGGGATGCCTAGAGGGTTGTTTGATCCTGTTTCGTGTAGAAATAGG
AGATGGAGGGTTGTTAGGGCTGTGATAATGAAGGGTAGGATAAAGTGGAAGGTAAAGAATCGTGTAAGGGTAGGGCTATCTACTGAGTAACCACCTC
AAACTCATTGGACTAGGTCTGTTCCGATGTACGGGATGGCGGATAGCAAGTTTGTGATTACTGTGGCTCCTCAGAAGGATATTTGGCCTCATGGGAG
GACATAGCCTATGAAGGCTGCTGCTATGGTTGTGAGTAGGNNNNNN

Go4_390          (chr4:125728606)
CTGACATGTTCATCTTCCACCCCTAGTACACAGTCAGAGACATCCTAGGCCTATTCGTCTTTCTCCTGACCCTGATAACATTAACACTATTCTCACC
AGACCTCCTAAGAGACCCAGACAACTACACTTTAGCCAACCCCCTAAACACCCACCCCACATCAAACCCGAATGATACTTCCTATTTGCCTACGCAA
TTCTCCGATCTGTCCCCATAAACTAAGAGGCGTCTAGGCCTATTATATCTATTCTCGTCCAGCAAAAATTCTATTCTCACAATACAAAAACAAAGAT
ATATCCCGCTCATNNNNNN

Go5_67           (chr5:142927566)
GTCCCTTGAACACCATCCTCCGTGAAATCAATATCCCGCACAAGAGTGCTACTCTCCTCGCTCCGGG

Go11_371         (chr11:83018072; chrX:19753155)
TAGTGTTCAGCACGTTATGCTTTCTACTGCAATTGACTCGTATTTTCTAACCTCGGTTACATCAATCCTGAAGTGGAGTATACGATTGATCTACATC
TTCATCACCATTGGAAAAGTAGCATTCGTACTATACTTATTCATGATCCTATTCATATTGCGAGTCACATCCCAAATCGAATATCAAATACTCACAT
GAGCCTGCTCTTGTAGTACAGAGCAATACACTAGTCTTGTAAACCGGAAACGAAGACCTCCTTCCAAGGTGATATTCAGAGAAGAAGTCCTCGACTC
CACCATCAGCACCCAAAGCTAACATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGAGACAAATTTGGGTGCCACC
```

```
Go15_391        (chr15:37175640
GGTCTGCACATCCCCATAAACAGATATGTTTGGTCCTGGCCTTTCTATTAGCTTTTAGTAAGATTACACACGCAGGCACCCCCGCCCCAGTGAAAAT
GCCCTCTAGATCACCCAGATCAAAAGGAGCAGGTATCAAGCATGCACAAATGCAGCTCAAAACACTTTGCTCAGCCACACCGCCAGGGGAAACAGCA
GTGATAAACTGTTAGTAATAAACGAAAGTTTAAGGTATACTGATATCTAGGGTTGGTCAATTTCGTGCCAGCCATCGTGGCCATACCATTAACCCAA
GTTAATAGAACTCGGCATAAAGAGTGTTTAAGGTCTGGCCCTCATAAAGCTAAACTCCATCTAAAGTGTAANAACCCTCAGCTGAATANATATACTA
TGAAAGTGGCTTTATACCTGAGACACATAGTTAGACCAACTGGATAGAACCACTAGCTT

Gcl39-2         (chr9:27204874)
TCCCTAATCGAAAACAAAATACTCAAATGAACCTGCCCCTGTAGTATAAGCTAATACACCAGTCTTGTAAACCGGAAANNNNN

Gcl39-1,5       (chr16:47290120)
ACCTCTTCACCCTAACCTGAATCGGAGGACAACCAGTAAGCTACCCCTTCATTACCATTGGGCAAGTAGCATCCGTACTATACTTCACGACAATCCT
ATTCCTGATACCAATCACATCCCTGATCGAAAACAAAATACTCAAATGAACCTGCCCTTGTAGTACAGACCAATACACCAGTCTTGTAAACCGGAAA
NNNNN

Gcl39-8         (chr11:25888029)
TACACACCAATACACCAGTCTTGTAAACCGGAAANNNNN

Gcl40-11a       (chr5:123203800)
CCCAACGAAATATCAACCGACCTGCCTCCTACAAAAGTACATAACACATAAGATCATTTATCGTACATAGCACANNNNN

Gcl39-7         (chr5:162567507)
CCCAACCTGAATCGGAGGACAACCAGTAAGCTACCCCTTCATTACCATTGGGCAAGTAGCATCCGTACTATACTTCACGACCATCCTGTTCCTAATG
CCAGTCACATCCCTAATCGAAAACAAAATACTCAAATGAACCTGCCCTTGTAGTACAGACCAATACACCAGTCTTGTAAACCGGAAACGAAGACCTC
CTTCCAAGGGCATATTCAGAGAAAAAGTCCTCGACTCCACCATCAGCACCCAAAGCTAATATTCTAATTTAAACTATTCTCTGTTCTTTCATGGGGA
GACAAATTTGGGTACCACCCAAGTATTAGCTAACCCATCAATAATTATCATGTATGTCGTGGC

Gcl47-3         (chr9:27211514)
CACCTCCCTAATCGAAAACAAAATACTCAAATGAACCTGCCCCTGTAGTATAAGCTAATACACCAGTCTTGTAAACCGGAAANNNNN
```

Appendix I. Alignment of mapped HVI numts and previous numt reports with high identity.

```
Numt1_1           CCAAGTATTAGCTAACCCACCAATAATTGTTATGTATGTCGTGCATTACTGCCAGTCACCATGAATAATATACAGTACTATAAACACCCAAATACCTATA
Rok8              ....................................................................................................
AY530149.1        ....................................................................................................
L76766.1          ....................................................................................................
AF240455.1        .................................................................................................G..
Muk5              .................................................................................................G..
AF240453.1        .....................................A...........................................................G..
AF250888.1        .................................................................................................G..
AF250889.1        ....................................................................................................
AY530150.1        ....................................................................................................
Muk9              ...............T.....................A...........................................................G..
AF240451.1        .................................................................................................G..
Muk8              .......................................................................................T.........G..
Rok7              .....................................A......................................................C....G..
Rok10             ....................................................................................................
AF250890.1        .....................................................................G...............G...........G..
AF250891.1        .....................................................................G..............................
muk7              ............................................................................................C....G..
AF240450.1        .....................................A......................................................C....G..
Rok5              .....................................A......................................................C....G..
AY530151.1        .....................................A......................................................C....G..
AY530152.1        ???????????????......................A......................................................C....G..
AY530153.1        .....................................R.....................................................?C....G..
AF240458.1        .....................................A......................................................C....G..
AF250887.1        .......................................................A.............G......................C....G..
Muk4              ...............T.....................A......................................................C....G..
AY530154.1        .....................................A......................................................C....G..

Numt1_1           ACACATACAAAACCCAACGAAATATCAACCGACCTACCCTCTACAAAAGTACATAACACATAAAATCATTTACCGTACATAGCACATTCCAGTTAAATCG
Rok8              ..........????......................................................................................
AY530149.1        ....................................................................................................
L76766.1          ....................................................................................................
AF240455.1        ....................................................................................................
Muk5              ..........????......................................................................................
AF240453.1        ....................................................................................................
AF250888.1        ...................................................................................................A
AF250889.1        ....................................................G..............................................A
AY530150.1        ....................................................G...............................................
Muk9              ..........????......................................................................................
AF240451.1        ...................................................................................................A
Muk8              ..........????.........................C.C..........................................................
Rok7              .T........????......................................................................................
Rok10             ..........????.....................................................................................A
AF250890.1        ................CG.......A..........................................................................
AF250891.1        .................G.........C...C...G................................................................
muk7              .T........????.........................C.C......................G..................................A
AF240450.1        .T.....................................C.C......................G..................................A
Rok5              .T........????.........................C.C......................G..................................A
AY530151.1        .T.....................................C.C......................G..................................A
AY530152.1        .T.....................................C.C......................G..................................A
AY530153.1        .T...............G.....................C.C......................G..................................A
AF240458.1        .T.....................................C.C......................G..................................A
AF250887.1        .................G.....................C.C......................G..................................A
Muk4              .T........????.........................C.C......................G..................................A
AY530154.1        .T.....................................C.C......................G.......?..........................A

Numt1_1           TCCTCGCCCCCATGGATGCCCCCCCTCAGATG
Rok8              ................................
AY530149.1        ................................
L76766.1          ................................
AF240455.1        ................................
Muk5              ................................
AF240453.1        ................................
AF250888.1        ................................
AF250889.1        ................................
AY530150.1        ..........................T.....
Muk9              ............................????
AF240451.1        ....T.......C...................
Muk8              ............................????
Rok7              ............................????
Rok10             ....T.......C..................A
AF250890.1        ................................
AF250891.1        ..............................G.
muk7              ............C...................
AF240450.1        ............C...................
Rok5              ............C...................
AY530151.1        ............C...................
AY530152.1        ............C...................
AY530153.1        ............C...................
AF240458.1        ....T.......C...................
AF250887.1        ............C...................
Muk4              ............C...................
AY530154.1        ....?.......C..................A


Numt2_1           CCAAGTATTAGCTAACCCATCAATAATTGTCATGTATGTCGTGCATTACTGCCAGCCACCATGAATAATGCACAGTACTACAAATGTCCAACCACCTGTA
AF240456.1        ....................................................................................................
```

```
AF240448.1          ...................C..........T........................T............................................
L76760.1            ....................................................................................C...T...........
AY530145.1          ............................A..........................A............................................

Numt2_1             ACACATACAACACCCAACGGAATACCAACCAATCCATCCCTCACAAAAAGTACATAACACATAAAGTCATTTATCGTACATAGCACATTCCAGTTAAATC
AF240456.1          .........................................................................C..........................
AF240448.1          ....................................................................................................
L76760.1            ......................................T.....................................C.......................
AY530145.1          ....................................................................................................

Numt2_1             ATCCTCGCCCCCACGGATGCCCCCCCTCAGATA
AF240456.1          ................................G
AF240448.1          .................................
L76760.1            G........................T...A...
AY530145.1          .................................


Gcl18-1             TGCCAGTCACCATGAATAATGCACAGTACTATAAACACCCAAACACCTGTAACACATACAACACCCAACGAAATATCAACCGACCTGCCTCCTACAAAAG
Muk4                ....................AT..............................T........????.....................A..C..C.......
Muk6                ....................AT..............................T........????.....................A..C..C.......
muk7                ....................AT..............................T........????.....................A..C..C.......
Rok5                ....................AT..............................T........????.....................A..C..C.......

Gcl18-1             TACATAACACATAAAGTCATTTATCGTACATAGCACATTCCAGTTAAATTACCCTCGTCCCCACGGATGCCCCCCCTCAGATG
Muk4                .......................C.........................C.T.....C.........................
Muk6                .......................C.........................C.T.....C.....................????
muk7                .......................C.........................C.T.....C.........................
Rok5                .......................C.........................C.T.....C.........................
```

VITA

Iván Darío Soto-Calderón was born in Medellín (Antioquia), Colombia on May the 3rd 1975. He received his B.Sc in 1999 from Universidad de Antioquia in Colombia, with emphasis in Zoology and Genetics. His undergraduate training included field work with manipulation and taxidermy of small mammals and the use of molecular tools in anthropological genetics of Native Colombians. He then enrolled in graduate school and focused on a project intended to integrate historical records and molecular data to reconstruct historical demographic processes of human isolates in northwest Colombia. He received his M.Sc. from the University of Antioquia in 2003 and decided to apply to graduate programs aiming for an opportunity in ecological and conservation genetics. In 2005, he was accepted as Ph.D. student in the Conservation Biology program at the University of New Orleans, where he collaborated with several projects of African duikers (genus *Cephalophus*) and Malagasy butterflies (*Heteropsis*) and developed his dissertation on the evolutionary dynamics of nuclear translocations of mtDNA (numts) in African Great Apes, Iván is back in Colombia, where he holds a position at the University of Antioquia and is currently developing his own research program in population genetics and conservation of Colombian fauna.